UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
SECÇÃO AUTÓNOMA DE HISTÓRIA E FILOSOFIA DAS CIÊNCIAS

THE POSSIBILITY OF FREE ACTION

DOUTORAMENTO EM HISTÓRIA E FILOSOFIA DAS CIÊNCIAS

MARIA JOANA TRIBOLET DE ABREU RIGATO

TESE ORIENTADA POR:

Professor Doutor António José Teiga Zilhão

e Professora Doutora Margarida Telo da Gama

2015

Documento especialmente elaborado para a obtenção do grau de Doutor

Para o Andrea, o Gabriel e a Clarinha

ACKNOWLEDGMENTS

These past five years have been quite a journey. I entered the academic arena and got to know the jittery buzz of discussing philosophy with the very best players in my field; I had to face many of my unexpected biases and unconfessed fears that literally took my sleep away for two years; I became a mother, twice. Throughout my endeavors, I was never alone, I never doubted that I would eventually reach the finish line of this investigation, and that is due to the generosity and friendship of many people whose help I can finally acknowledge here.

First of all, I thank my advisor, Professor António Zilhão, whose patient and (very!) demanding mentoring have enabled me to produce this piece of work. He has believed in me since the very beginning of my philosophy studies 17 years ago and I feel extremely privileged for having had the opportunity to learn as much as I have from him. Despite our sometimes difficult discussions, I do hope our relationship of mutual respect will remain fruitful in the years to come.

Second of all, I thank Professor Margarida Telo da Gama, an excellent physicist, for having so generously accepted the unexpected task of supervising a philosophy thesis. I enjoyed every bit of our interdisciplinary work together.

I also thank Professor Rui Moreira, who accompanied the scientific side of my work during the first years of this research, as well as professor José Croca, who was the first person in the Center for Philosophy of Sciences (CFCUL) I turned to in 2010, and who opened the door for me. Professor Olga Pombo, the director of CFCUL, was the one to welcome me in and to provide me with all the conditions I could have asked for in order to develop my work. I truly appreciate all her help over these years.

I met many fine people at CFCUL and at the Faculty of Sciences in general, but two good friends must be mentioned in particular: Gil Santos and João Cordovil, with whom I learned a great deal and who were always willing to help me along the way. Some parts

of this dissertation benefited directly from some of our discussions and from their honest comments on earlier drafts.

The possibility to attend amazing international meetings about Free Will enriched my work enormously. Many people made it possible for these trips to happen, in particular: Thomas Müller, Rani Lill Anjum and Robin Hendry, to whom I am truly indebted. I thank all the philosophers I had the chance to talk to and have discussions with on those occasions, especially: Robert Kane, Timothy O'Connor, Randolph Clarke, Neil Levy, Laura Ekstrom, Derk Pereboom, Joshua Shepherd, Christopher Franklin and Adina Roskies. Their generosity and warmth was incredible. A friend once told me that he thought the Metaphysics of Free Will was the best field in Philosophy to work in, because it had the nicest philosophers. I am not familiar with many other fields, but from what I have experienced during these past few years, he might well be right.

Over the course of these years, I had to take three formal examinations and had the privilege to have three very good philosophers as my external examiners: Rui Sampaio, Robert C. Bishop and Joseph Berkovitz. I thank all of them for their precious comments on my work. I am especially indebted to Robert C. Bishop, who kindly accepted my invitation to come present his work at our center without us ever having met before, and who was generous enough to return the following year. His work had a strong influence on my research and the endless discussions we engaged in were truly precious. As is his friendship.

I also thank the people I met along the way who have been struggling with Free Will, Emergence or Conscience just as I have and whom it has been a real pleasure to encounter in conferences and workshops and to keep in touch with by email. It is a pity we are so scattered around the globe; I do wish we had the chance to meet more often. I am very grateful to all those who have been kind enough to read portions of my work, to discuss it with me or to help me with the prose here and there (many of those mentioned above and also Olivier Sartenaer, Pranab Das, Brian Keeley, Rui Costa, Simon Kittle and William Murphy): I hope you won't feel disappointed with the final product.

Last but not least, I must thank my newfound community at the Champalimaud Center for the Unknown. In particular, I thank Alex Gomez-Marin, who welcomed me first at the CCU, and Zachary Mainen, the director of the Champalimaud Neuroscience Programme, a brilliant neuroscientist who fortunately has managed to keep his interest in philosophy alive and who has been so generous as to work with me in the intersecting points of our two fields of research.

Finally, I am also very thankful to the Portuguese Foundation for Science and Technology (FCT) who provided the financial support without which this investigation would not have been possible.

And now, to my dearest friends and family…

Agradeço, antes de mais, à minha mãe, irmãos Paulo e Miguel e sobrinho André pelas infindáveis explicações de matemática e física nos primeiros tempos do doutoramento. Apesar daquele auspicioso regresso às ciências duras, o meu caminho acabou por se manter longe das equações. Mas a satisfação intelectual que os estudos que fiz com a vossa ajuda me proporcionaram valeu mesmo a pena, como correr a maratona depois de 17 anos de sedentarismo.

Agradeço ainda ao Paulo pelo acolhimento caloroso em Oxford, por todas a conversas que tivemos acerca destes temas e pela ajuda na formatação final da monografia. Ao meu irmão Tiago, às minhas queridas cunhadas e a todos os meus sobrinhos, um abraço apertadíssimo por estarem sempre a torcer por mim.

Obrigada à comunidade do Graal, pela riqueza espiritual que me equilibra. Aos amigos, pela partilha de vida que me enche.

Aos meus pais Benedita e António, vai a gratidão profunda por todo o apoio afetivo e material com que me têm acompanhado ao longo da vida. Nestes últimos anos, a vossa ajuda com os netos tem sido essencial para o meu trabalho, e o vosso amor fundamental para a minha estabilidade emocional. Esta tese deve-vos muito.

Ringrazio inoltre i miei carissimi suoceri Irene e Paolo che sono la generosità in persona e una presenza essenziale nella vita della mia famiglia. È in gran parte grazie al vostro sostegno, pazienza e inarrestabile gioventù che questo lavoro è arrivato a compimento.

E eis que quase faltam as palavras para as três pessoas centrais da minha vida: o meu marido Andrea, companheiro da verdadeira aventura e dos verdadeiros mistérios, e os meus filhos Gabriel e Clara, fonte de luz, alegria, criatividade e confiança no futuro. A nossa casa é um “urban campfire” construído e reavivado a cada dia. Por toda a comunhão, obrigada. Ainda que os caminhos desta tese por vezes me tenham embaciado os olhos, os vossos não falham nunca.

CONTENTS

ABSTRACT

The present dissertation develops an agent-causal libertarian theory of action and free will. The backbone of its argumentative structure is that 1) there can be no agency without an agent-cause; 2) there can be no agent-causation without indeterminism; 3) hence, libertarianism is the best option for any realist view about action.

The first step in this argument is a defense of agent-causalism. I develop a taxonomy of behavior and agency and argue that both libertarian and compatibilist event-causal accounts fail to provide an adequate description of the differences we find between non-actional behaviors and full-blooded actions.

Agent-causal accounts, however, are usually met with suspicion because of their requirement of an irreducible agent with downward causal powers. My second step aims to respond to this concern by presenting a scientifically informed account of emergence as a way to show that the natural order of the world is compatible with the existence of irreducible and causally effective entities, such as the agent's self. I defend the thesis that natural supervenience does not have to be challenged by this possibility, as downward causation requires only the break of causal closure and bottom-level indeterminism. I argue that both these conditions are unproblematic.

The third step is the contention that we have many reasons to believe that consciousness is an emergent property. Moreover, the unity of phenomenal experience suggests the existence of a unified self as the bearer of conscious properties. The conscious self is the irreducible substance-cause who exercises its causal power over the alternatives left open by the probabilistic laws governing its neural substrate. When the conscious self intervenes, agency happens. When it is passive, bodily movement reduces to mere behavior.

Given that the requirement for fundamental indeterminism renders my account of agency an incompatibilist account of free will, the final step of my dissertation is the assessment of the classical objections against libertarianism. After analyzing the most important arguments for and against contemporary views akin to my own, I respond to all the

objections and conclude that agent-causal libertarianism is the most plausible and satisfactory view of how actions are possible and free.

Key words: Action, Free Will, Emergence, Consciousness, Self

RESUMO

Na presente monografia é desenvolvida uma tese incompatibilista acerca da possibilidade da ação num mundo governado por leis físicas que escapam ao controlo do agente. Segundo esta tese, se as relações de causa-efeito no universo natural forem estritamente deterministas, não só não são possíveis ações livres, como não é possível sequer haver ações.

Na Introdução, é apresentado o debate clássico em torno do livre-arbítrio, a sua história e o estado da arte do problema no contexto da tradição analítica. Salienta-se ainda a importância das neurociências para o renovado vigor do debate na atualidade, nomeadamente devido à pressão que o conhecimento detalhado de como certas propriedades mentais dependem de funções neuronais exerce sobre posições libertistas que queiram reivindicar para o agente uma dose elevada de autonomia.

No segundo capítulo, intitulado "*Setting the stage for agent-causation*" (o porquê de uma teoria da causação pelo agente), o ponto de partida é a tese de que toda a ação pressupõe um *self* a partir do qual o agente se relaciona com o mundo e dele se distingue. Em seguida, essa tese é desenvolvida através de uma taxonomia do comportamento humano, em que são distinguidos vários níveis de intervenção do agente na sequência causal que produz a ação e, consequentemente, no controlo que aquele exerce sobre esta. Há muitos comportamentos humanos que não podem ser considerados ações na medida em que a ligação causal entre certos acontecimentos neuronais e o movimento corporal não é mediado por nenhuma intenção explícita do agente. Mas há também comportamentos intencionais (como os pequenos furtos de um cleptomaníaco, por exemplo) que, embora sejam motivados por razões do agente (suas crenças e desejos), não são produzidos por ele. Em todos estes casos, o agente não é o autor do comportamento, daí este não poder ser considerado uma ação, apesar de aparentar sê-lo.

Esta visão da ação tem dois pressupostos teóricos: primeiro, uma visão da causação como sendo uma relação entre uma substância e um acontecimento (*substance-causation*), e

não uma relação entre acontecimentos (*event-causation*); segundo, uma abordagem não reducionista do agente, na medida em que o poder causal deste sobre os movimentos do seu corpo não pode ser equivalente à soma dos efeitos causais das partes que o constituem.

A conceção da ação que aqui se desenvolve identifica-se com a corrente denominada, em inglês, *agent-causalism* (o que, em português, se pode traduzir por teoria da causação pelo agente). Ao contrário do que normalmente é tido como certo, a teoria da causação pelo agente não é uma versão do libertismo. É uma posição segundo a qual uma ação é causada diretamente pelo agente enquanto substância e não pelos seus estados e acontecimentos mentais. Como tal, pode ser assumida independentemente da filiação teórica que se tenha em relação à questão do livre-arbítrio.

O terceiro capítulo, intitulado "*Irreducibility in nature*" (a irredutibilidade na natureza), trata do tema da emergência. A motivação para tal reside na necessidade de desenvolver uma visão do agente enquanto substância irredutível causadora da ação, que não só faça jus às exigências da teoria da causação pelo agente apresentada no primeiro capítulo, mas seja simultaneamente compatível com a imagem que a ciência contemporânea nos dá do mundo natural. Para que a ideia de que o *self* do agente emerge naturalmente do cérebro sem coincidir com este seja razoável, há pois que baseá-la numa teoria da emergência suficientemente coerente e plausível, teoria essa que se procura desenvolver neste capítulo.

Começa-se por rever a história do debate em torno do conceito de emergência e por apresentar uma sua definição. Passa-se em seguida à análise de alguns dos casos de putativa emergência que são frequentemente citados na literatura da especialidade e entre os físicos da matéria condensada (que se opõem a uma visão excessivamente reducionista da realidade), após o que se conclui que as teorias físicas que tratam dos vários níveis de organização da matéria (mecânica quântica, mecânica estatística, termodinâmica, mecânica clássica, etc.), não estabelecem entre si uma relação de continuidade. Mostra-se que há fenómenos macro (ex. temperatura) cuja descrição tem inevitavelmente de recorrer a modelos que são irredutíveis às teorias que descrevem os fenómenos micro subjacentes, o que revela situações de emergência epistémica. No

entanto, a ciência física não tem ferramentas que lhe permitam partir da emergência epistémica e daí inferir conclusões acerca de uma suposta emergência ontológica. Não obstante a precaução que se impõe, defende-se que o emergentista teórico pode afirmar que, ao contrário daquilo que muitas vezes se supõe, a sua visão da estrutura da realidade é compatível com a física atual.

De seguida, este capítulo dedica-se a estabelecer as condições de possibilidade da emergência ontológica. Dado que se opta por preservar a relação de sobreveniência natural segundo a qual toda a mudança mental assenta numa mudança neuronal subjacente, as condições de possibilidade da emergência terão de ser as seguintes: 1) a quebra do princípio do fechamento causal do mundo físico e, 2) indeterminismo fundamental. Enquanto a primeira condição é consensual entre todos os defensores da emergência, a necessidade da segunda para a plausibilidade da causalidade descendente é referida por muito poucos autores e a sua fundamentação é um contributo original desta tese. A argumentação apresentada leva à conclusão de que não é possível haver causalidade descendente (efeitos causais da entidade emergente sobre o substrato físico que lhe deu origem e a sustém) sem que ambas estas condições estejam garantidas, e mostra-se que ambas são compatíveis com o conhecimento científico de que dispomos atualmente.

No quarto capítulo, intitulado "*The conscious self*" (o self consciente), passa-se à questão de saber se algum âmbito da realidade nos oferecerá evidência de que a emergência ontológica é, não só uma possibilidade, mas um fenómeno efetivamente existente. É sugerido que a experiência fenomenológica é uma parte da realidade que é simultaneamente inegável e distinta da base fisiológica que a origina. As propriedades conscientes têm qualidades que tornam impossível a sua redução ao corpo físico, suscetível de uma descrição em termos de estrutura e função. Por essa razão, tais propriedades apresentam-se como boas candidatas a fenómenos ontologicamente emergentes.

Por outro lado, a experiência consciente de um sujeito apresenta-se-lhe como intrinsecamente unificada e como emanando de um único ponto de vista. Segundo o argumento da unidade da consciência, esse ponto de vista não pode ser a soma de pontos

de vista parciais, pelo que as propriedades conscientes requerem a existência de um sujeito unificado que as contenha. Esse sujeito é o *self* irredutível cuja existência a teoria da causação da ação pelo agente postula.

O quinto e último capítulo, intitulado "*Free Will and alternative possibilities*" (livre-arbítrio e possibilidades alternativas), aborda os argumentos clássicos que têm sido debatidos em torno do problema da compatibilidade entre o livre-arbítrio e o determinismo, por um lado, e o livre-arbítrio e o indeterminismo, por outro. Dada a exigência, apresentada no terceiro capítulo, de que a estrutura subjacente a uma entidade emergente não seja governada por leis exclusivamente deterministas, por forma a que haja espaço lógico para o exercício do poder causal da dita entidade, a tese defendida nesta monografia acabou por se revelar incompatibilista. Portanto, torna-se pertinente confrontar as objeções que têm sido colocadas a uma tese deste tipo ao longo da história do debate em causa, apresentando também os principais argumentos usados em defesa do incompatibilismo em geral e do libertismo em particular, contra as posições compatibilistas. Nesse confronto, verifica-se que o incompatibilismo sai vencedor e que a teoria da causação pelo agente permite inclusive solucionar alguns dos impasses teóricos que se apresentam ao longo do caminho.

No final, é possível concluir que o libertismo *agent-causalist* aqui proposto é a posição mais capaz de conciliar, por um lado, a experiência fenomenológica da agência e a vantagem prática de conservarmos a nossa corrente classificação dos comportamentos humanos consoante a capacidade de autoria do agente, e, por outro, as constrições provenientes do conhecimento científico que retiram às posições dualistas tradicionais qualquer plausibilidade. Ao contrário destas, a ideia do *self* consciente como entidade emergente, associada a uma visão indeterminista do funcionamento neuronal, é compatível com uma descrição naturalista do ser humano e da sua interação com o mundo. O homem é um sistema biológico regulado por leis físicas, mas é simultaneamente detentor de uma capacidade de tomar decisões e intervir no mundo através do seu corpo, cujo espaço de atuação é garantido pelas possibilidades alternativas deixadas em aberto pelo indeterminismo subjacente. O *self* do agente é um ator a pleno

título no desenrolar da história do mundo, sem que isso implique qualquer quebra das leis da natureza.

Palavras-chave: Ação, Livre-arbítrio, Emergência, Consciência, Self

“To deny the reality or logical significance of what we can never describe or understand is the crudest form of cognitive dissonance.”

(Thomas Nagel)

1. INTRODUCTION

> "The central question in philosophy at the beginning of the twenty-first century is how to give an account of ourselves as apparently conscious, mindful, free, rational, speaking, social, and political agents in a world that science tells us consists entirely of mindless, meaningless, physical particles. Who are we, and how do we fit into the rest of the world?"
>
> (John Searle, 2004)

I have taught Philosophy in high school for several years. The Philosophy curriculum in Portugal includes material relating to the topic of Action for tenth-grade teachers to teach in six 90-minute lessons. The problem of the compatibility between Free Will and Determinism is only one part of this topic, which means that I, as a teacher, was required to cover the Free Will debate in no more than four lessons. However, given the nature of the subject matter, I never managed to get through it in fewer than six or even eight lessons. I was concerned that the issue, which I found terribly disturbing at an existential level, would not be adequately explored if covered in fewer lessons.

Whenever I sensed that my students were perhaps not sufficiently confused or overwhelmed by our discussions, I would try to employ better texts and thought experiments in order to get them to realize just how intricate the subject is and how, as a result, their self-image as free and responsible agents was being challenged.

I am not sure whether this excessive compulsion of mine had a positive effect on my students, but I do know that the concern I felt during those discussions has had a lot to do with bringing me to where I am today. Over the past four-and-a-half years I have struggled with most of the philosophers of our time who have discussed the free will problem and

tackled its subtleties. I have striven to find my own way out of these conundrums and do actually believe that I have succeeded to some extent.

In this Introduction, I will present briefly the traditional problem of Free Will and the main strands of its current debate. I will then explain the reasons why I believe the problem is more exciting now than ever and why it is intricately connected with many other fields in and outside Philosophy. Finally, I will present an overview of the structure of the present dissertation and suggest that the account that is developed here is a new and valuable contribution for the debate.

1.1. The problem

The idea we have of ourselves as free agents is based on our introspective awareness of having certain beliefs and desires, of forming intentions to act according to those belief-desire pairs, and from our having the vivid impression that those intentions cause us to behave in a certain way. We have the feeling we are in charge of what we do most of the time. Despite obvious constraints that prevent us from having the ability to do everything we might want to do (humans cannot fly, for example), we are still given a limited number of alternatives and we can choose between them (even though I cannot fly, I can still decide to either take the stairs or the elevator to go down from the third to the first floor). Since we endorse our decisions and perceive ourselves as the causes of our actions, we accept the responsibility for their consequences. If I tell a friend some information about the schedule of the bus she must catch in order to arrive at a meeting on time and that information is wrong and causes her to miss the bus and her meeting, I feel sorry for my mistake and apologize. I believe that several occurrences in the world and in my friend's life happened as they did because of me and therefore I assume my responsibility for having caused them. So more so when I intended for those consequences to happen. We all believe our actions can actually cause changes in the course of events and that is what most stimulates us in our everyday endeavors.

The problem of free will arises when our knowledge of how the laws of nature determine the course of events in the world makes us suspect that we are not the true source of our actions, nor the source of our will. Spinoza put it brilliantly:

> "For instance, a stone receives from the impulsion of an external cause, a certain quantity of motion, by virtue of which it continues to move. (…) Further conceive, I beg, that a stone, while continuing in motion, should be capable of thinking and knowing, that it is endeavoring, as far as it can, to continue to move. Such a stone, being conscious merely of its own endeavor (…) would think that it continued in motion solely because of its own wish. This is that human freedom, which all boast that they possess, and which consists solely in the fact, that men are conscious of their own desire, but are ignorant of the causes whereby that desire has been determined."[1]

Spinoza believed men are just as constrained in their actions as this stone is in its movement and hence their freedom is just as illusory. Spinoza was an *incompatibilist* about free will: according to him, free will is incompatible with determinism. If universal determinism is true, then the complete state of the world at each instant is caused by the complete state of the world at the instant that immediately preceded it, in such a way that no other state could have occurred. The terrible implication of this for human action and free will is that the complete state of my body (which includes my neurons and their inter-relations, my memory and personality) determines what my wishes, values, opinions and criteria for choosing will be every time I have to make a decision. I have no alternative possibilities of action: given the person I am, plus my past, all the details of my present circumstance and the laws of nature, I cannot choose nor act otherwise.

Incompatibilism has two horns: Spinoza's view, which falls under the category we now call *nihilism* or *hard determinism*, and its direct rival position, which is called *libertarianism*. According to libertarianism, men do possess free will, so the causal processes that are relevant for the production of free actions must somehow escape deterministic laws.

Many philosophers in the past thought the above description of a deterministic world in which free will would be a mere illusion just could not be true, it could not be the whole

[1] Spinoza, B. (1674), p.390.

story. In order to defend libertarianism, they opted for a dualist account of human agents. If people are not composed just of a physical body, but have also a deep self or a soul that somehow escapes the rigid constrictions of deterministic laws, then free will can be saved. René Descartes[2] and Immanuel Kant[3], for example, defended this type of account.

However, dualist accounts encountered many problems: on the one hand, the question of how could there be an interaction between an immaterial substance and the physical world; on the other, the inconvenience of postulating the existence of a supra-natural substance, thus challenging Occam's razor and the reductive tendencies of modern science (especially since the advent of genetics and molecular biology). These problems rendered this position very unpopular.

Nevertheless, the advent of quantum mechanics in the first quarter of the twentieth century seemed to open new possibilities for libertarianism. Nature was shown to obey probabilistic laws at the most elementary level and determinism was called into question. Maybe libertarians did not have to postulate an immaterial soul in order to endow the agent with alternative possibilities of action after all. Despite the suspicion of some[4] who claimed that randomness would diminish, rather than enhance, the agent's control, authors like Roderick Chisholm[5] and, more recently, Timothy O'Connor[6] opted for a view called agent-causalism which grounds libertarianism on the agent's substantial causal powers to indeterministically cause her actions. This view was supposed to carry with it less controversial metaphysical assumptions than Cartesian or Kantian alternatives, without giving in to nihilism about free will.

By the end of the twentieth century, more naturalistic forms of libertarianism were put forward by philosophers such as Robert Kane[7] and Laura Ekstrom[8]. Their goal was to do

[2] Descartes, R. (1649, 1664).
[3] Kant, I. (1781, 1788).
[4] Cf. Smart, J.J.C. (1961).
[5] Chisholm, R. (1964).
[6] O'Connor, T. (2000).
[7] Kane, R. (1998).
[8] Ekstrom, L.W. (2000).

away with any "extra-factors" that science would not explain, including any mysteriously irreducible agent-cause, and to ground the agent's control on the causal connection between her reasons, her intentions and her actions. Other authors, like Carl Ginet[9], tried to develop accounts of free actions that would simply regard them as uncaused events which would nevertheless be non-random, as they would be explainable in terms of reasons and purposes.

All these accounts have problems and face numerous objections, stemming from both empirical and theoretical grounds. Some objections are aimed at the specific details of each account, while others are directed at the libertarian stance as a whole. The main criticisms are the following. On the one hand, there is a strong suspicion regarding the possibility that a large and warm system such as the brain may be sensitive to quantum fluctuations. Even if it were, it seems implausible that indeterminism should be located precisely where it is most convenient (at the moment of choice or immediately before, depending on the specifications of different accounts), and never at times when its presence would remove the agent's control altogether (as for instance between the formation of an intention to act and the overt action). On the other hand, many have considered that if an agent can act in one way and in another under the exact same circumstances, then her action cannot be appropriately explained and thus will be unintelligible or even irrational. According to this objection, libertarian free will is an incoherent concept and therefore impossible, regardless of the truth or falsity of indeterminism.

Because of all the problems each and every one of its versions faced, libertarianism remained quite unpopular throughout the twentieth century. Apart from its persistent defenders and some open nihilists such as Ted Honderich[10], Derk Pereboom[11] and Saul Smilansky[12], most philosophers took the compatibilist stance and argued for the

[9] Ginet, C. (1990).

[10] Honderich, T. (1993).

[11] Pereboom, D. (2001).

[12] Smilansky, S. (2002).

possibility of free will in a deterministic world (or a world in which all the relevant structures such as the neural network work deterministically).

Compatibilism has its roots in the view, defended by Thomas Hobbes[13] and David Hume[14], of freedom as the absence of constraint. According to this standpoint, the type of free will people are interested in is that which can found responsibility ascriptions, and that is simply the ability to act according to one's will, without being subject to compulsion nor constraint. During the second half of the twentieth century, however, this position was re-elaborated in response to increasingly complex formalizations of the arguments in favor of incompatibilism[15], which motivated compatibilists to embrace different argumentative strategies: Some, headed by Peter Strawson[16], chose to dissociate metaphysical questions regarding free will from the psychological phenomenon of responsibility attribution, and focused on the necessary and sufficient conditions for the latter. Others, led by Harry Frankfurt[17], called into question the idea that actions could be free only if the agent could have acted otherwise (also known as the Principle of Alternative Possibilities or PAP). Others still, consented to the truth of PAP but focused their attention on questioning the incompatibilist inference from the truth of determinism to the impossibility of acting otherwise.

As usual in philosophy, no position is uncontroversial. Libertarians are accused of defending a solution that is either implausible or incoherent, or both. Free will nihilism is considered to be an unattainable position as it contradicts our basic experience of agency and would have dangerous implications for society, in what regards responsibility ascriptions and justice. Also, given our current scientific view about the world (after the quantum revolution of the mid-1920s), determinism is much less certain, which gives us no reason to embrace a view that relies on its truth. Compatibilist accounts of free will are considered to provide too weak a concept of free will and, as Kant noted, to fail in

[13] Cf. Chappell, V. (1999).

[14] Hume, D. (1748).

[15] Most notably, the Consequence Argument (see section 5.2.1).

[16] Strawson, P.F. (1962).

[17] Frankfurt, H. (1969).

their attempt to solve "with a little quibbling about words, that difficult problem on the solution of which millennia have worked in vain"[18].

Despite the ongoing debate, by the end of the twentieth century, we seem to have reached a stalemate.

1.2. New challenges

In the last decades, astounding developments in the sciences of the brain have given this dispute new vigor. The possibility of identifying the neural correlates of mental processes such as intending and deciding opened empirical pathways for approaching the problem of human freedom. Benjamin Libet's experiments[19] were a corner stone in the now called neurophilosophy of free will. They allegedly showed that our conscious intentions to act are a relatively late part of the decision-making process, which is caused and initiated at an unconscious level. If Libet's results and their philosophical interpretations should prove to be true, they would have devastating consequences for the libertarian, and arguably even for the compatibilist. In fact, it would be very difficult to maintain that an unconscious decision is something over which the agent can have a satisfactory degree of control.

Libet's experiments, as well as the numerous follow-ups[20] that purported to confirm its results with increasing accuracy, have been long criticized for several reasons[21]. Above all, philosophers have complained that the neural signals that neuroscientists measure in these kind of experiments need not be interpreted as the unconscious process of making a decision or forming an intention to act. Evidence that the brain's activity preceding the

[18] Kant (1788), 5: 96–97.

[19] Libet, B., et al (1983).

[20] Cf. Soon C.S., Brass M., Heinze H.-J., Haynes J.-D. (2008) and Fried I., Mukamel R., Kreiman G. (2011).

[21] See, most notably, Alfred Mele's numerous treatments of the subject: 2006, 2009, 2014a, 2014b.

conscious intention to act is predictive of the ensuing action to some degree is not surprising if one takes, as most contemporary libertarians do, mental events to have a very close relationship with neural events. A conscious intention is part of an ongoing flux of mental and neural activity, and it is only natural that some sort of neural preparation precedes it and that it will have a higher probability of meeting our unconscious biases. Thus, the bold inferences some authors have made from the data gathered in Libet-type experiments (which typically deal with spontaneous decisions between indifferent options preceded by no deliberation) to the claim that conscious will is an illusion in any situation[22] have been met with great suspicion.

Regardless of the actual relevance of Libet's results, the debate on these experiments and their philosophical implications has been useful, as it showed how close we have come to having empirical results providing definitive answers to long lasting questions regarding free will. Neuroscience is reshaping the philosophical discussion and there is no turning back. The idea that the mind supervenes on the brain is corroborated every day by hundreds of studies which gather increasingly detailed information about the neural substrate of mental processes. We cannot discuss the free will problem without addressing also the mind-brain relationship and, in Wittgenstein's words, the problem of identifying "what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm"[23].

Experiments with animal models throw light on the workings of the human nervous system and, more importantly, they are explanatorily efficient even when it comes to human behavior[24]. For example, spontaneous actions like the ones performed in Libet tasks can be successfully modelled in mice, and the data gathered in experiments done with these small mammals has recently been claimed to have important implications for the dispute over the correct interpretation of the brain activity measured in humans performing those tasks[25]. More than ever, we have evidence that biological continuity is

[22] Cf. Wegner, D. (2002).

[23] Wittgenstein, L. (1953), p.161.

[24] Cf. Gold, J.I., Shadlen, M.N. (2007) and Glimcher, P. (2005).

[25] Cf. Murakami, M. et al. (2014) and Rigato, J. et al. (2015).

the criterion with which to understand human beings as part of nature, to unveil the mechanisms of their animal bodies and to interpret their psychological lives. But is there any room for free will if human agents are nothing more than animal agents with the added upgrade of superior cognitive functions such as language and abstract reasoning?

Some have embraced this challenge without renouncing to a robust concept of free will. According to Helen Steward[26], for example, all actions are inconsistent with determinism because no action can fail to be up to the agent (and nothing can be up to the agent if things are fully settled by the past and the laws of nature). Nevertheless, according to her, human agency does not require some mysterious break in the evolutionary continuity we find everywhere else in nature. In fact, "the concept of agency is an outgrowth of the concept of animacy" and it applies "unproblematically to many animals"[27], not just to humans.

I am very sympathetic to this view. People are animals and their behavior is biologically constrained to a great degree. Nevertheless, there is an incredible variability in nature that scientists classify as noise (i.e. data that their models do not capture) and which on my account is the product of the animal's agential intervention in an underdetermined course of events. Free actions come in degrees; they can be very simple in mice, and highly complex in humans, but the agent as the ultimate cause is present throughout the spectrum. This dissertation is an attempt to show how this agential intervention may work in the natural world.

1.3. Overview

The novelty of my account of free action and free will lies in its argumentative structure. I defend an agent-causal libertarian theory, but instead of basing it on a priori arguments for a certain concept of free will, I ground it in the empirical assessment of how human

[26] Steward, H. (2014).

[27] *Idem*, p.xi.

behavior can be classified as actional and sub-actional, together with an analysis of the requirements for, and implications of, this distinction. The backbone of my thesis is that there can be no agency without an agent-cause, and no agent-causation without indeterminism. As my readers can probably guess, I am hoping this will provide an extra argument in favor of libertarianism.

The present dissertation is structured in four steps. First, I will contend that agent-causalism is required by any realist view about action. In order to show this, I will develop a taxonomy of behavior and agency and claim, on the basis of the "disappearing agent argument", that event-causal accounts, both libertarian and compatibilists, fail to provide an adequate description of the differences we find between non-actional behaviors and full-blooded actions.

Agent-causal accounts, however, postulate the existence of an irreducible agent with downward causal powers, which is usually regarded as an implausible requirement. In order to assess this contention, I will develop a scientifically informed account of ontological emergence as a way to show that the natural order of the world is compatible with the existence of irreducible entities. I define ontological emergence as a relation between entities (substances or properties) belonging to different hierarchical levels, in which the higher-level entity is simultaneously dependent on and autonomous from the lower-level entity or structure. Emergent entities possess novel capacities whereby they can have causal effects that cannot be explained solely on the basis of the causal powers of the lower-level entities. Given the requirement of natural supervenience, which I will contend should not be abandoned easily, these causal effects must be exercised downwardly. I will claim that usual objections against emergence can be adequately answered if two conditions are met: the break of causal closure and bottom-level indeterminism. I will show that the requirement that all causes are fundamental causes is unwarranted and that it is plausible to think that laws of nature leave more than one alternative open in order for the emergent entity to affect the course of events. Given these two conditions, downward causation becomes possible without any risk of overdetermination nor epiphenomenalism.

After having assessed the theoretical possibility of ontological emergence, I will proceed to discuss its actual existence. I will argue that even if one assumes a cautious attitude regarding the actual presence of emergent properties in the material world (the world that can be described, measured and explained by science), we have many reasons to believe that consciousness is exceptional. First-person descriptions of phenomenal experience (the only ones we have) are irreducible to third-person descriptions of the physical substrate that produces it. After having argued for this, I will contend that the unity of phenomenal experience suggests the existence of a unified self as the bearer of conscious properties – that same self that we call "agent" in the context of action production.

Given the requirement for basic indeterminism, my account, which started off as merely agent-causalist, has become an incompatibilist account. Hence, an assessment of the main arguments that have been put forward in the traditional debate about the compatibility between free will and determinism is in order. In the last chapter of this dissertation, I will present the famous Consequence Argument, the multiple Frankfurt-type examples and the more recent manipulation cases and will conclude that incompatibilism is the best account of free will, not only because of the need for an irreducible agent, which fosters the need for indeterminism, but also given these more classical arguments. Finally, I will present what I consider to be the two main objections against libertarianism, which are based on the idea that indeterminism is inimical to control, insofar as it renders the agent's decision uncertain and unexplainable. I will respond to these objections and conclude that agent-causal libertarianism is a plausible and satisfactory view of how actions are possible and free, despite our physical nature as animals subject to natural laws.

2. SETTING THE STAGE FOR AGENT-CAUSATION

> "When animal agents exist in a world, the unfolding of that world through time must wait upon decisions and choices which have to be made by those animals – not just temporally (…) but also metaphysically."
>
> (Helen Steward, 2012)

2.1. Actions and agents

In neuroscience, action is defined in opposition to response[28]. There is a continuum at the farthest end of which we find simple reflexes (immediate and automatic motor responses), while at the other extreme end lie voluntary actions (not directly determined by any external stimulus).

> "In contrast to responses, actions are behaviours where it is either impossible to find an eliciting stimulus or where the latency and/or magnitude of the behaviour vary so widely, that the term 'response' becomes useless."[29]

The idea of a voluntary action in neuroscience has to do with this freedom from immediacy[30] by which an animal agent (human or not) makes a decision or self-initiates a spontaneous action (like a boy jumping up in the air unexpectedly, while playing[31]) in the absence of clues from the environment that might serve as evidence in favor of it over any alternatives. This implies the demarcation of an agent's self, as the source of the spontaneous actions, the controller of controlled processes and, in the case of humans,

[28] Cf. Brembs, B. (2011), Haggard, P. (2008).

[29] Brembs, B. (2011), p.933.

[30] Cf. Shadlen, M.N., Gold, J.I. (2004).

[31] This example is used by Brembs (2011), p.935.

as the mediator between the animal's body and the social system[32]. The agent's recognition of herself as an entity distinct from the environment and from the others, and her identification with that part of the physical world that she can control to some degree – her body – is a precondition for any voluntary action to take place, as well as for it to be interpreted as such by others.

However, the interdependence between action and the sense of selfhood is a two-way entailment, for the demarcation of the agent's self can happen only if the agent acts in the first place. An animal can distinguish its *self* from the world via a mechanism called *reafference*[33], by which it can naturally and unconsciously tell apart those sensory stimuli that are consequences of its own actions and thereby are under its control (e.g. the darkness caused by its eye blinks), from those that are not. So, just as "in order to understand actions, it is necessary to introduce the term self", "the concept of self necessarily follows from the insight that animals and humans initiate behaviour by themselves" [34].

Similarly, in philosophy, despite the enormous controversy over the definition of every concept involved in a theory of action and the relation obtaining between them, the distinction between what *happens* to people and what people *do* is foundational to most accounts. While an infinity of events take place at each instant in the universe (the radioactive decay of an atom, the falling of a leaf, the arrival of a photon at my retina), only a subset of these can be classified as actions: the ones that are made intentionally by an agent, or which at least are the involuntary side-effect of an intentional act (e.g. to misread something counts as an action because, in doing it, the agent intended to read)[35].

For an act to be intentional is for it to be the execution of a plan, which is the mental representation of the future action to be performed. The plan (which might go from very simple to highly complex) is the representational content of the intention. Previous to the

[32] Cf. Baumeister, R.F. (2010).

[33] von Holst, E., Mittelstaedt, H. (1950).

[34] Brembs, B. (2011), p.936.

[35] Davidson famously defended the idea that to act is to intentionally do *something* in his "Agency" [in Davidson, D. (1980), essay 3].

conception of that plan, of course, the agent must have certain reasons to do something or to act in such a way as to achieve a certain goal. Reasons are usually taken to be pairs of desires and beliefs. Thus, the desire to drink plus the belief that water is in the fridge and that the enactment of certain behaviors is necessary if I am eventually to swallow it, will lead me to form a certain plan and eventually to execute it, opening the fridge, taking out the water, filling up a glass with it and drinking it. Note that representational attitudes such as intending, believing or desiring have both a certain content and a certain orientation[36]. What makes the attitude I have towards this specific content an intention is that its orientation is that of an *executive attitude*, a disposition to put my plan into practice[37].

To make a decision to A is to actively form an intention to A (leaving aside the alternative of not A-ing). The intention will thus be the product of the decision that triggers it. This does not imply that every action has to be the result of a decision. More often than not, there is no need to decide, for there is no uncertainty about what to do (e.g. whether to eat or not when I am hungry, there is food in front of me and there are no reasons not to eat it). In such cases agents act on intentions that derive from their standing preferences (that arise out of habit or are the result of intentions already formed). Likewise, not all decisions are the result of a deliberative process (the process of considering the reasons in favor of each possible alternative), for there are impulsive decisions made on the spot that are nevertheless the settling of previous conditions of uncertainty and bring about an intention to act.

To sum up, an action, as the word indicates, is an event of which the agent is the *active* promoter, not a passive pawn in the game. Sometimes (but not often) the English language helps us see this difference: the *rising* of an agent's arm is an event (a motion that took place in a certain place at a certain time) that could have been caused by a myriad of factors, such as strings pulling it up, like a puppet arm. But if that event is caused

[36] Cf. Searle, J.R. (1983).

[37] Cf. Mele, A. (2009), pp.3-7.

by the agent's intention to bring it about, then her *raising* her arm intentionally is an action that the agent performed[38].

It is assumed by any action description that the agent is some*one* distinct from all others and that she has a self that is both the subject of the decisions and intentions which led to the action, as well as the subject of the action itself: "*I* opened the door" is an action; "the door opened" is an event. I am the subject of the action I perform in the first case, whereas the door is not the subject of any action in the second; it is the subject only of a proposition that describes an event that just happened to take place without it having been *done* by anyone.

The difference between acts that simply involve the agent as its locus ("my arm rose up") and actions that the agent endorses as her own ("I raised my arm") is foundational to the phenomenological experience of agency as well. Pathological conditions such as the Anarchic Hand Syndrome or the Tourette Syndrome are paradigmatic of situations in which the agent's self is exceptionally not part of the causal process that leads to what would otherwise appear to be an action, even though her body is. In cases of anarchic hand syndrome, for instance, patients perform involuntary movements with their hands and make comments such as "it will not do what I want it to do" or "it does what it wants to" [39]. In Tourette's cases, the agent's "body is animated by a continuous stream of urges that demand specific and often complex oral and motor responses"[40], which leads to a never-ending succession of physical and vocal tics, and sometimes cursing. The involuntary bodily movements are so frequent and intense, that it can become difficult to tell apart the patient's self from the phenomena directly caused by the disease: "I consist of tics—there is nothing else" [41].

[38] Cf. Helen Steward (2014, pp.33-35), who makes a distinction between bodily movements (events) and bodily *movings* (actions), based on Jennifer Hornsby's analysis (1980, p.3).

[39] Pacherie, E. (2007), p.212.

[40] Buckser, A. (2008), p.167.

[41] Sachs, O. (1998), p.98.

Helen Steward has recently proposed an interesting definition of the agent as "an entity that has a body and can make that body move in various ways"[42]. According to her, the owner/body distinction that arises when a certain degree of complexity in the biological hierarchy of species has been reached is the basis for action, defined as an execution of the agent's power of self-movement.

> "Most animals of any appreciable degree of complexity (…) are possessors of a capacity for a kind of top-down determination of what will occur with respect to the movement of their own bodies, in such a way that their contribution does amount to something over and above the contribution of the processes inside them which eventuate in the resulting bodily movements."[43]

The thesis I will develop in this dissertation will be similar in many aspects to Steward's view. I too believe that, for there to be an action, there must be a form of top-down determination by the agent herself, of the strictly physical events taking place in her body. This means that the agent's causal power is somehow irreducible to the causal powers of her parts, including those of her propositional attitudes, such as intending or deciding. Not only is mental causation an ability the possession of which is a pre-condition for an agent to be considered such, but agent-causation is a fundamental form of intervention is the world, one which grounds the very possibility of agency.

However, Steward considers that, even though all actions are settlings of matters[44] and, because of that, they all require top-down causation, not all of them are intentional. For instance, absent-mindedly scratching one's head counts as an action, but a "sub-intentional" one[45]. This is a clear point of disagreement between us, since I follow Anscombe and Davidson's view instead, according to which all actions are intentional under some description, and thus prefer to consider such behaviors as to unconsciously

[42] Steward, H. (2014), p.32.

[43] *Idem*, pp.16-17.

[44] "The core idea at the heart of this notion of settling a matter is that of a question that is capable of being resolved in different ways at all times up until a certain moment – the moment of settling – at which point something that happens causes it to become resolved in one particular way" [Steward, H. (2014), p.39].

[45] Cf. Steward, H. (2014), p.34.

jiggle one's foot as events that are the product of the activity of the organism's body but not something that an agent did[46]. Even if we do not see eye to eye on the extension of the set of actional behaviors, Steward and I agree, however, that all actions require a form of downward causation and that for there to be such a downward causal link, two conditions must be assured: first, the physical world cannot be causally closed (which is quite consensual among those who defend downward causation); second, there must be indeterminism at the bottom-most level of organization of matter (a condition that appears to be much more controversial). I will address both these conditions in chapter 3. In the remainder of the present chapter, I will develop my account of action as an agent-caused type of behavior.

2.2. A taxonomy of behaviors and actions

The cases of missing agency that I mentioned before (Anarchic Hand and Tourette syndromes) are just two among an endless list of cases that span across a whole spectrum of behaviors. Agency, I believe, comes in degrees, and understanding what changes from case to case along the spectrum will allow us to understand how the different elements involved in action production come into play (Table 1).

Let us start with the Anarchic Hand syndrome (AHS). People suffering from this disease find themselves totally unable to control the movements of one of their limbs, which engage in behavior that seems goal-directed and often elicited by inputs from the environment, but which is unintended. The "alien hand"[47] will unbutton a shirt that the

[46] The question of distinguishing an entirely automatic behaviour such as breathing, from a "semi-automatic" one such as scratching one's head is of course a tricky one. I will address it in more detail in the next section.

[47] I am using the term "alien hand" here in order to distinguish the hand that is functioning independently from the patient's will, from the one that is still under the patient's control. However, it is important to note that efforts have been made by experts in the past years to distinguish the Anarchic Hand Syndrome from the Alien Hand Syndrome, which is a condition in which patients feel that one of their uppers limbs does not belong to them – whereas in the disorder that is concerning us here, there is no such feeling [Cf. Pacherie, E. (2007)].

patient keeps trying to rebutton with the other hand, it will slap the patient in the face, it will refuse to cooperate in tasks such as cooking or reading the newspaper.

TYPE OF BEHAVIOR/ACTION	DESCRIPTION	EXAMPLE	CAUSAL STRUCTURE	DEGREE OF CONTROL
zombie behavior	Bodily mechanics without consciousness	Anarchic hand syndrome	Stimulus → response	cannot be stopped nor controlled
		Absent-minded behavior		
alienated behavior	Implanted agency	Utilization behavior	Urge → response	
		Manipulation	"Alien intention" → movement	
		Hypnosis		
reactive behavior	Bodily mechanics with consciousness and no control	Tourette's (tic without premonitory urge)	Neural input → response	
		Nervous assassination/climbing	Emotion → physical reaction	
purposive behavior	Bodily mechanics with consciousness and urges	Tourette's (tic with premonitory urge)	Urge → release	can be stopped and controlled effortfully
		Kleptomania		
		Addiction		
		Anger	Emotion → movement	
→ THE AGENT ENTERS THE PICTURE				
spontaneous action	Automatic reasons-responsive acting	Routines	Stimulus + standing intentions → response	can be stopped and controlled for reasons
		Self-regulation habits		
action-on-the-spot	Acting upon fast decisions	Libet's decisions	Urge → decision → movement	
		Immediate decisions	Habit/tendencies → decision → movement	
deliberative action	Acting upon decisions made through deliberation	Rational decision	Deliberation → decision → movement	

Table 1

"I'd light a cigarette, balance it on an ashtray, and then my left hand would reach forward and stub it out. It would take things out of my handbag and I wouldn't realize so I would walk away. I lost a lot of things before I realized what was going on."[48]

[48] Karen Byrne, a patient suffering from this syndrome after a surgery to cure epilepsy in which her brain's *corpus callosum* was cut (article by Dr Michael Mosley for *BBC News Health* 01.20.2011).

AHS patients retain the ability to act purposefully with the rest of their body and they feel alienated from the behavior of the anarchic limb which, despite belonging to them, seems to have "a mind of its own"[49]. Their phenomenological experience is that of having a part of their own body behaving in a way that is not intended and cannot be inhibited, which they try to refrain (for example, by blocking it with the other hand) and about which they express frustration (they complain that the hand is "always trying to get into the act"[50]).

Complex as it might be, this type of behavior is like a reflex, a purely mechanical response to a certain input (internal or external), with no conscious mental state involved. A more common situation that shares the main characteristics of AHS is automatic behavior. Imagine those everyday situations in which people answer questions they are posed without realizing what they are saying (when they are watching television, for instance), or when people pick up an object (their keys or glasses, say) and absentmindedly put it somewhere else without realizing what they are doing – which might make them have to look for that object for twenty minutes the next time they have to leave the house. When asked why they put the object in the place where it was eventually found, people will say they do not know. They do not even remember putting it there, just like people suffering for AHS do not know what their hand is doing if they do not see it doing it. The type of behavior that is typical of all these cases (AHS, absentmindedly saying or doing something that we cannot explain afterwards) is what I call ***zombie-like behavior***. Needless to say, it does not consist in actions at all.

Another interesting condition in which people lose the capacity to inhibit stimulus-driven behavior is called Utilization Behavior (UB). Patients that suffer from this disease are dependent on external stimuli in such a way that they cannot act unless solicited and their perception of an object is taken as an "order" to use it[51]. If they see a pair of glasses, they will immediately put them on, and if a second pair is shown to them before the first pair has been taken off, they will put that second pair of glasses over the first pair; if a hammer

[49] *Idem*, p.212.

[50] *Ibidem*.

[51] Lhermitte, F., Pillon, B., Serdaru M. (1986).

and a nail are put before them, they will immediately hammer the nail to a nearby wall, independently of the appropriateness of the context, or lack thereof.

Strangely, though, unlike patients suffering from AHS, UB patients do not seem surprised by their behavior and if asked why they did those things, they will give evasive explanations such as that "they thought they were duties that had to be carried out and that they were natural things to do"[52]. This means that a very significant difference between UB and AHS is that patients suffering from the former do not seem to realize the inadequacy of their behavior and so they endorse it as if it were an action they performed voluntarily. Interestingly, this difference in the phenomenology of agency is reflected also in the inability of UB patients to act purposefully in the absence of cues from the environment (they exhibit apathy when not externally stimulated).

Also, UB patients do not explain their actions in terms of their own intentions or desires, which suggests that they are not moved by endogenous motives. They are impelled to use objects by a general sense of duty. Together with the apathy patients exhibit, this has led authors to hypothesize that the structures that are impaired in cases of UB somehow involve the general capacity for "agentive self-awareness", which prevents patients both from engaging in spontaneous actions and from realizing that their automatic behavior was not autonomously brought about by their own intentions[53]. When patients recover from the lesions in their frontal lobes that correlate with the disease, their actions regain independence from the environment, and they express perplexity at their previous behavior and at "the fact that they had no controlling thoughts of their own"[54].

I believe this is analogous to what one might imagine could happen in cases of manipulation, such as those that philosophers like to imagine and which I will discuss further ahead (cf. sections 5.2.2. and 5.2.3.). If an evil Dr. Black were to manipulate my mind and brain in such a way that at the sound of midnight on my clock I would spontaneously feel the urge to kill my next door neighbor, my acting on that urge would

[52] Pacherie, E. (2007), p.212.

[53] *Idem*, p.216.

[54] Lhermitte, F., Pillon, B., Serdaru M. (1986), p.332.

seem to be intentional, it would apparently be the product of my mental states. However, since that urge would have been caused by Dr. Black's manipulation, the ensuing intention to kill would not be truly mine, as it would have been implanted in me by another. A more realistic and frequent case is deep hypnosis. Subjects that are hypnotized report the experience of an "absorbed and sustained focus of attention on one or few targets", a "relative absence of judging, monitoring, and censoring", and the feeling that one's own responses are "automatic (i.e., without deliberation and/or effort)"[55]. All these elements seem to adequately describe what we can infer from UB patients' reports of their experience. The objects that they encounter solicit in them an immediate response, that is not subjectively felt as a reflex, but as a voluntary need to use the object, as if there was nothing else one could do in that situation. As hypnotized subjects describe, there is a "sense of automaticity wherein thinking is no longer felt as preceding action but action is felt as preceding thought"[56].

These sorts of cases are different from the zombie ones in that relevant mental states are no longer absent from the agent's conscious field. Patients feel that they are acting and that they are doing what they want to do. However, common intuition will exempt these patients from responsibility in case their actions lead to undesirable consequences. If I end up killing my next door neighbor and in court some brilliant lawyer finds a way to prove that Dr. Black manipulated me, I will probably walk free because I *did not know what I was doing*. The same goes for someone suffering from UB or a hypnotized person, should they end up hurting someone else when they engage in their compelled behavior. Legally and morally, they can hardly be considered imputable, for their specific condition (the disease or the hypnosis) is what effectively produced their putative actions. Once they "wake up" from that condition (either because they are healed or because the effect of hypnosis wears off), they realize what they have done and become mortified.

Are ordinary people, as well as legal systems, wrong in considering that these actions are abnormal and that the agents should not be blamed (nor credited, for that matter) for

[55] Rainville, P., Price, D.D. (2003), p.111.

[56] *Idem*, p.113.

their consequences, in spite of the fact that during the act they feel that they are doing what they want to do? I do not think so. The mental states that led to the action were not originated in the agent's selves, they were implanted in them by an alien element: the brain damage, the manipulation, the hypnosis. There was no agential intervention because the agent was somehow dormant, passive, incapable of "judging, monitoring, and censoring". The process that led to the action was like a stimulus-response process, even if a conscious mental state was perceived to be the cause (an urge, an intention, a sense of duty, depending on the situation at stake), instead of a neural unconscious state, as in the Anarchic Hand cases. But such a conscious mental state was not a mental state of the agent; it just highjacked the agent's mind for a certain period of time, like a virus that uses the host cell's machinery in order to replicate. Given the particular nature of these cases, I call the type of behavior that they describe ***alienated behavior***. And again, I consider it to be sub-actional.

Let us consider now the type of cases that I call ***reactive behavior***. They regard those situations in which our own conscious mental states are perceived to be amongst the causes of our behavior but not with the conscious consent of our control system. The famous "deviant causal chains" associated with cases such as Davidson's nervous climber[57] are quintessential examples of such situations. Imagine that an agent wants to perform a certain action (to shoot someone, for instance) but, due to her nervousness, her body performs the action for her (the involuntary twitching of the trigger-finger causes the gun to fire) without it having been *her* to decide or intend to do it then. The agent's emotions caused her behavior reactively, through a causal chain of which she was not in charge, even though she could see the whole process happening "inside" herself.

This is also what happens to Tourette's patients, most of the time. A patient suffering from Tourette Syndrome (TS) is always self-aware and lives a fruitful life, with a job, a family and friends. However, her condition makes it very difficult (often impossible) for her to control the outburst of motor and vocal tics, some quite simple such as barking or eye blinking, others much more complex, such as punching herself, touching objects or

[57] Cf. Davidson, D. (1973).

people, bending and twitching her body, uttering inappropriate sentences (coprolalia) or repeating what other people say (echolalia). The disease makes her also more prone to many other behavioral symptoms such as obsessive thoughts (TS is very often associated with Obsessive Compulsive Disorder), negative reactions to novel situations due to anxiety, great difficulty in inhibiting impulsive behavior, episodes of incontrollable rage, etc. The symptoms vary a lot according to the context, depending on whether the subject is alone or with other people, in the intimacy of her home or in a public place, whether she is concentrated in a certain task in which the flow of her behavior can follow a smoothing rhythm or if nothing in particular is catching her attention.

When a patient's tics are involuntary and uncontrollable, her behavior is reactive, like the twitching of the nervous assassin's finger. In those cases, tics just have to be released, like a sneeze. That does not happen only with physical movements. Frequently, TS patients find it impossible to go through a written text because they feel the uncontrollable need to "read each line many times, (…) to line up each paragraph to get all four corners symmetrically in [their] visual field, (…) to 'balance' syllables and words, (…) to 'symmetrize' the punctuation in [their] mind, (…) to check the frequency of a given letter"[58], etc. Also in cases like these, which resemble the compulsive episodes typical of anxiety disorders, the patient's mental undertakings should not be considered to be actions; they would better be classified as forms of reactive behavior.

However, Tourette's is such a complex disease, with so many degrees and variations, that it can fit into different levels of behavior/agency. While growing up, most TS patients start experiencing premonitory sensory phenomena which might allow them to sense that a certain tic is about to arise and to prevent it occasionally, due to some training. This is not easy: even when it is possible to prevent the tics, that prevention costs the patient a lot of effort, it increases stress and it can only last for a short period of time[59]. Nevertheless,

[58] Sachs, O. (1995), p.86.

[59] Cf. Banaschewski, T., Woerner, W., Rothenberger, A. (2003); Buckser, A. (2008).

the possibility of intentionally blocking the tic opens the door for the behavior of a Touretter to be much more voluntary than what *prima facie* might appear[60].

When the TS patient gives in to the tic, she feels that she is actually *doing* it, that she is not assisting passively at the event of the tic coming to be. Imagine Anne is a girl talking to a friend while constantly balancing her glasses, patting her friend in the shoulder and pressing her foot against the ground in circular movements all around the spot where she is standing. All these movements could eventually be stopped, were she to make a strong effort to block them. However, she chooses not to because she feels at ease with her friend who has grown used to her tics and she might as well store her self-control for occasions on which the tics might be much more disturbing or inappropriate. Anne finds it very difficult to inhibit her tics, but not impossible. And certainly she endorses the tics as something that she does, with varying degrees of awareness and control, depending on the circumstances.

When Anne lets her tics be released with no effort to control them, she undertakes a somehow hybrid behavior that I call ***purposive***, which takes place when conscious mental states related to desires (urges, impulses, needs) are the direct causes of the behavior, without there yet being the control of the agent via her intention to do what she does. In these situations, there is a mixture of voluntary and involuntary elements in the agent's performance: the voluntary removal of the inhibitory breaks that might block the tic versus the involuntary outburst of the specific tic that comes about. When on the contrary Anne blocks her tics, she becomes the helmswoman of her ship – her body – steering it according to her decisions, in spite of the highly conditioned elbow room at her disposal. In those cases, the transition from sub-actional behavior to a full-blooded action is accomplished.

But one might as well ask why is it that I consider the voluntary tic that a TS patient might endorse to be something distinct from the action that the same patient performs when she successfully tries to control her tic. There are several reasons that justify this analysis: the first is phenomenological. Tourette's patients themselves perceive their tics as

[60] Schroeder, T. (2005)

primitive urges that, due to the disease, they find it much harder to inhibit than "normal" people[61]. Tics are like a biological impulse that gives them no satisfaction, but to which they must surrender, sooner or later, in one form or another.

> "What is uncontrolled about a tic is not the movement itself, but the need to move, the urge to make a very specific gesture or sound. (…) It means that tics can be displaced—they can be delayed or relocated to times and places where others will not observe them."[62]

When patients prevent the tic, they do so for conscious reasons that they decide to act on, such as the need to avoid the social disadvantages of the tic, like being stared at and misjudged by other people. Let us consider some real life cases of such a "displacement":

> "One woman, growing up on a farm, took several long walks daily in the woods. Her family attributed these to a solitary or soulful nature; in reality, she told me, she simply needed a place where she could release her tics, which she had to suppress in the house. She would walk for miles, 'twitching and spitting like a maniac,' then return home unsuspected."[63]

> "In a situation of close social contact, where satisfying the urge for a facial or neck tic would be very noticeable, almost all of [the] informants [of a study on TS patients] said that they would occasionally induce a tic in the leg or foot instead. By performing that tic intensively—clenching the toes hard or hyperextending the ankle, for example—they could divert energy away from the facial tic, and perhaps suppress it altogether."[64]

The strategies these Tourettic patients found in order to remain unobserved are clearly intentional and rationally motivated. Their tics, on the contrary, are irrational impulses dictated by the neurological disorder that they would rather avoid. The difference between the phenomenological experiences under these two circumstances must be

[61] Cf. Bennett, the surgeon presented to us by Oliver Sachs (1995), described his syndrome as "a disease of disinhibition" (p.84).

[62] Buckser, A. (2008), p.175.

[63] *Ibidem*.

[64] *Idem*, p.176.

accounted for and it seems to me that the best explanation for this difference must be that tics and displacement strategies are two radically different types of events.

The second reason for putting these types of behavior under two different categories is based on the intuitions that guide our common practices of responsibility attribution. For example, one of the problems that parents with children suffering from Tourette's face is the difficulty of diagnosing this disease, which is of the utmost importance given the behavioral symptoms that are associated with the syndrome. After the diagnose is made, the episodes of misconduct of the child can be interpreted not as a behavioral or emotional issue, but rather as direct consequences of an abnormal neurological condition. Since TS is considered a disability, students suffering from this impairment are fit for independent education programs, according with the legislation of each country, and "may not be punished or disciplined for behaviors that are caused by or are a manifestation of their disabilities"[65]. This legal protection agrees with the common intuition by which we naturally tend to excuse someone who might behave inappropriately once we learn that that person has Tourette's, even if we understand that her behavior was voluntary, to some degree. If a TS patient insults someone out of a vocal tic (giving in to the urge to shout "fatso" at an obese person, for instance), she does not do it in order to meet an object of desire (the pleasure of causing discomfort, say), like a malicious person would. She does it to escape the growing tension and discomfort that her urge is causing her, as if it were a growing itch that can be alleviated only by scratching. The urge that moves her is ultimately the outcome of a neural process devoid of any propositional content, while the control she exercises over it is an action made by reasons. Those reasons are of course supervenient on their neural correlates and might even be identical to them (this is something we will discuss later), but they have a cognitive content and express the agent's character in a way the primitive urges uninhibited by Tourette's do not. Like Timothy Schroeder put it:

[65] United States' section 504 of the Rehabilitation Act of 1973.

> "The way in which a Tourettic individual resists an urge to tic says much about the quality of the individual's will, but the urge itself says nothing."[66]

Impulsive control disorders like kleptomania, as well as substance abuse, also fit into this category of purposive behavior. They too are characterized by the failure to resist a harmful impulse that one would rather not have, a failure which is caused by a neural mechanism that is much more difficult to oppose and much less expressive of the agent's self than the common psychological mechanism of *akrasia* or weakness of will[67].

Drug addiction is defined as "the loss of control over the intense urges to take the drug even at the expense of adverse consequences"[68]. The degree to which that control is effectively lost is still under some controversy[69] and it is an empirical fact that the effects drugs have on each person, both at the neural and the behavioral level, are very diverse. Some people manage to go through years of recreational consumption without ever becoming addicted while others get trapped in the net; many addicts manage to quit drugs without treatment while many others do not. Both genetic and social factors have been proven to be extremely influential, but still the degree to which each person retains the possibility to choose whether to use drugs or not on each occasion seems very hard to determine.

The mainstream view[70] is that drug addiction is a chronic disease, caused by the highjacking of several brain circuits related to reward, motivation, learning, inhibitory control and executive function. Due to the disease, patients lose the ability to value any sources of pleasure besides the drug (their dopamine D2 receptors, which are responsible for the sense of reward, are reduced) and their behavior becomes compelled, either by the need to escape the physical symptoms of withdrawal, which are extremely painful, or

[66] Schroeder, T. (2005), p.119.

[67] Cases of weakness of will are those everyday cases in which, in spite of being under no compulsion, a person does not act in accordance with her judgment of what is best to do under the circumstances (e.g. when someone who is on a diet eats a chocolate cake she herself judged better not to eat).

[68] Volkow, N., Li, T.-K. (2005), p.1429.

[69] Cf. for example, Kennett, J. (2013) and Satel, S., Lilienfeld, S.O. (2013).

[70] Cf. "Drugs, Brains, and Behavior: The Science of Addiction" by Dr. Nora Volkow, in the National Institute on Drug Abuse (NIDA) website: www. drugabuse.gov.

by the psychological mechanisms of craving. Also, growing evidence has shown that drug addicts are less capable of acting on reasons reflecting long-term goals than non-addicts, as well as more reactive to drug-related stimuli catching their attention and triggering automatic motor mechanisms of response.

> "[D]rugs can trigger bottom-up, involuntary signals originating from the amygdala that modulate, bias, or even hijack the goal-driven cognitive resources that are needed for the normal operation of the reflective system and for exercising the willpower to resist drugs."[71]

Once a person becomes addicted, drug-related thoughts become impossible to eliminate, while the drug craving makes the desire to take the drug the only motivation that the addict can experience. Unlike the pleasure given by the drug, which diminishes over time, the craving remains extremely strong even after long periods of abstinence and it is not experienced as the reasonable desire for something pleasurable, but as an intense feeling of "wanting" that is irresponsive to reasons.

Probably the objectors to the brain-disease model are right when they claim that, since most drug-addicts do eventually quit on their own (most of them do so before their 30's), to say that patients are powerless is neither truthful nor fair (not to mention harmful[72]). Many addicts just choose not to exercise the power they still have to fight their compulsion and they are responsible for that, just like the ones that do make the effort and succeed should be credited for it[73].

However, these considerations do not change the fact that, *when* hardcore addicts act out of compulsion, they are responding to an urge that is not an expression of their evaluative

[71] Noël, X. Van Der Linden, M., Bechara, A. (2006), p.30.

[72] Even though this has not been proven in the case of drug abuse, the belief in the disease model of alcoholism has been shown to be one of the primary factors in the likelihood of relapse in cases of patients treated for their drinking problems [Miller, W.R., Westerberg, V.S., Harris, R.J., Tonigan, J.S. (1996)]. Also in other cases, such as obesity, for example, evidence has shown that presenting people with the idea that they are prey to a disease they cannot control will make them feel powerless and make worse choices (like eating a worse diet) than subjects with the same disorder but who have not been "labeled" the same way [Hoyt, C.L., Burnette, J.L., Auster-Gussman, L. (2014)].

[73] Cf. Satel, S., Lilienfeld, S.O. (2013).

system. Their behavior is purposeful as they are doing what at that moment they most want to do and they can make small choices about when and how to use the drug – just like a TS patient can, when she displaces her ticcing in order to render it more discrete. Nevertheless, such behaviors should not yet be considered to be actions, given that the subject is not doing what she had previously decided and intended to do. Here, unlike in regular cases of weak-willed actions, her capacity for self-control has been severely compromised. The tendency for automatic action given the monopolization of attention by drug-related stimuli makes the option of resisting temptation so effortful that the alternative of just throwing the towel and taking the drug becomes almost irresistible – again, just like a Touretter, when she finally cannot control her ticcing anymore and an outburst of motor and vocal movements is released.

The major difference, of course, is that Touretters are born with a neurological disease that is part of who they are and can never be cured. Drug addicts, on the contrary, have made past choices that led to the neuroadaptations caused by drug use and, most importantly, they can improve greatly their condition through abstinence (even if, according to the brain-disease model, there is no possibility of their brain ever recovering completely). However, despite these major distinctions, at the moment of giving up their agential power in favor of their urges, addicts and Touretters are both expressing similar purposive behavior.

Kleptomania is similar in many aspects to the aforementioned disorders, even though it has been comparatively less studied and hence its neurobiology is yet poorly understood. It is defined as the "recurrent failure to resist impulses to steal objects that are not needed for personal use or for their monetary value" (and in which the stealing is not committed in response to hallucinations nor can it be better accounted for by other disorders, such as Antisocial Personality Disorder, Maniac Episodes, etc.)[74]. It is currently considered to be a chronic disease, with exacerbations and remissions[75], the phenomenology of which

[74] *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. (American Psychiatric Association, 2000).

[75] Cf. Grant, J.E., Kim, S. W (2002), p.378.

is clearly similar to what addicts experience during craving and substance use[76], as well as to the Tourettic need to release the urge to tic: stealing episodes are characterized by the experiencing of an "increasing sense of tension immediately before committing the theft" as well as "pleasure, gratification, or release at the time of committing the theft"[77] and the periods of voluntary abstinence are characterized by increasing urges. The kleptomaniac behavior is impulsive, repetitive and expresses an impaired inhibition that patients resent. They do not steal for personal gain or fun, they do it for "symptomatic relief"[78], and most importantly for our case, they experience shame and guilt afterwards[79], which may lead in some cases to considering the possibility of suicide, "to stop themselves from stealing"[80]. As in the case of TS patients' tics, kleptomaniacs do not understand why they steal the particular items they do, which is revealing of the fact that their reasons are not what moves them to the type of behavior performed.

I believe stealing is not an action in the case of kleptomaniacs, just like taking drugs or releasing a previously sensed tic in the two aforementioned examples were not. However, the behavior of a kleptomaniac is clearly purposive, as the patient has to articulate different levels of attention and movement in order to do something that is dangerous and shameful but that needs to be done to soothe the increasing tension she feels. She knows what she is doing and she does what she most strongly wants to at that moment, even though she would rather not want to do it – as the persisting efforts kleptomaniacs do to resist their urges confirm[81].

Harry Frankfurt famously argued for a hierarchical view of the human person as someone who is capable of having second-order volitions, that is, "capable of wanting to be different, in [her] preferences and purposes, from what [she] is"[82]. This is clearly an

[76] Cf. Grant, J.E., Odlaug, B. L., Kim, S. W. (2010).

[77] *Diagnostic and Statistical Manual of Mental Disorders*. 4th ed. (American Psychiatric Association, 2000).

[78] Grant, J.E. (2006), p.82.

[79] Cf. Grant, J.E., Kim, S. W (2002).

[80] *Idem*, p.380.

[81] *Idem*, p.381.

[82] Frankfurt, H. (1971), p.12.

element that is present from the level of complexity of purposive behavior on, and that was absent before. In cases of zombie, alienated or reactive behavior, there were not yet first-order desires or urges about which the person in question might have an approving or disapproving perspective. The subject could be annoyed at her behavior, as in cases of Anarchic Hand Syndrome, but the object of her disengagement was a mere physical movement, not a psychological feature that she disliked about herself. In contrast, in cases of purposive behavior, there is an urge that moves the person at the level of desire, that makes her *want* to do something, and that wanting is what is perceived by her as the motor of her endeavors. There are degrees inside this level, though. The tic that a TS patient chooses to release when nobody else is around is likely to be more primitive and uncontrollable than the act of stealing performed by a kleptomaniac. The involvement of mental states in the causal process that leads to the behavior is probably variable, the Tourettic urge being somewhat more physical (like an itch), the urge to steal being more emotional (like the sexual impulse when one is in love), and the craving for the drug being something in between, depending on the circumstances (on the type of drug, the level of addiction, the phase in which the addict is, etc.).

According to Frankfurt's analysis, someone who does not have second-order volitions, i.e., someone who does not care about her will, is not a person. He calls that being a "wanton" (and he includes in this category nonhuman animals and very young children), which I believe is a very adequate name to label patients that suffer from Utilization Behavior and act out of spontaneous desires that they do not question. Interestingly, Frankfurt uses the example of three different addicts to illustrate better the relationship between his concept of a person as a hierarchically willed being and his theory of the freedom of the will as the ability to want what one wants to want. I have not yet come to the point where I wish to discuss the concept of free will (I will come back to Frankfurt when I do that), but these addicts are useful for the previous analysis of degrees of agency that I am undertaking here.

Frankfurt distinguishes an unwilling addict (who has the second-order volition that taking the drug would not be – as it is – his first-order desire), a wanton addict (who has no second-order volition) and a willing addict (who is happy with his desire to take the drug

and would not have it any other way). According to Frankfurt, only the unwilling and the willing addicts are persons, and only the latter is free. The aspect I believe it is interesting to select from these three cases is whether any of them concerns a type of behavior that admits being considered to be an action. The physiological conditions are, by hypothesis, the same in all of them, and all three "succumb inevitably to their periodic desires for the drug to which they are addicted"[83]. So how would they fit the different categories into which I have been dividing human behavior? The answer is that all of them belong to the same category of purposive behavior. The fact that the wanton addict has no upper level perspective on his desires, or that the willing and unwilling addicts have opposite second-order volitions does not change the most important element of the situation, according to my analysis: the direct causal link between their wanting and their consumption without the intermediate intervention of their intention-forming power. Even if the willing addict intends to take the drug, it is not that intention that moves him, but his addiction, since, according to Frankfurt's scenario, "his desire to take the drug will be effective regardless of whether or not he wants this desire to constitute his will"[84]. The process of bringing about the behavior is produced by this addiction, which is much more similar to the Tourettic need to tic than to the process by which the unwilling addict might decide not to take the drug, for instance. The fact that one's second-order volition is coherent with one's first-order desire is secondary relative to the primary factor of the agent's intention not being what determines the action.

There is one last example of purposive behavior that might help us see where the frontier lies between a non-actional behavior and a proper action: very strong emotions. When hot-tempered people are furious and engage in a discussion, they often say and do things they might regret and that they had explicitly promised themselves not to say or do. That is why we often avoid situations that might get out of hand, like choosing not to talk personally to people that make us "lose our temper". These are everyday situations in which it is not a drink or a drug that take hold of our reactions, it is not a disease either, it is our emotions. In certain situations, they make us lose rational control over our actions

[83] *Idem*, p.17.
[84] *Idem*, p.25.

and sometimes say "it wasn't me, it was my resentment [or rage, or fear or jealousy] speaking"[85].

When instead we manage to do what we intend or decide to do, we act. Just like non-actional behaviors, actions too are diverse and can be classified according to the degree to which the agent contributes to them. The actions in which the agent is less involved are ***spontaneous actions***, such as routinely preparing one's breakfast. These are actions that are impelled by the agent's previously formed intentions but which do not require her full attention at the moment when they are made. When we go down the stairs, cross the street and enter the car, we are not consciously deciding to do all these things because we act out of "a will already formed"[86], or according to a "standing intention"[87]. These actions are under our conscious control nonetheless, and we do them for reasons. If these reasons change (e.g. if our car is parked somewhere else, or if we realize that we forgot something important at home and have to go up the stairs again, etc.), our actions will change accordingly. This is called *reasons-responsiveness* and it is a quality the essential core of all the behaviors I analyzed before lacks. Even if a drug addict can choose how much heroin he wants to shoot up, he is not deciding to take the drug *simpliciter*. Even if he is given all the reasons in the world for not shooting (all the bad consequences, the possibility of dying, of being incarcerated, of losing custody of his children), once he is addicted and all the severe circuitry changes have taken place in his brain, he has lost control over his decisions and his behavior will very hardly change. If it does, then it is not a mere purposive behavior anymore, it has become a proper action.

Another example of spontaneous decisions are self-regulation habits, such as not saying everything that comes to our mind if it is not socially adequate. Since childhood we have been taught these habits and any healthy person will train them gradually while growing up until they become natural and almost effortless. This does not mean that they are not under the agent's control. They are, and that is why one can decide to give in to one's

[85] A similar example is suggested by Velleman in his (1992), p.465.
[86] Kane, R. (1998), p.78
[87] Mele, A. (2009), pp.3-4.

impulses sometimes and feel "childish". Also, that control is precisely what one loses in emotionally charged situations such as a fight, when these inhibitions are removed.

It is important to note that actions such as these are common to nonhuman animals as well. They too act intentionally when they move in space in search for food and they too are capable of great self-control, with training and reinforcement.

> "If you want to see control in the wild, watch a predator stalking prey. A fox slowly, silently creeps up on the pheasant until she is close enough to spring with a good chance of success. Her pups watch and learn. When a pup first tries for himself, he is apt to spring too soon. Feeling the disappointment of losing out on food, he learns to bide his time. He learns to control his impulse to spring *now*."[88]

The main new element in these actions, both in human and non-human animals, is the agent. Certainly, these actions might seem like an automatic response to stimuli from the environment, according to a previously programmed algorithm, just like the utilization behavior of a patient who will hammer a nail to the wall once both these objects are put in front of him. The similarity derives from the fact that routine actions do not follow explicit deliberation, nor are they the outcome of a decision, since there is no uncertainty about what to do under the specific circumstances the agent is in. If I prepare my breakfast, I can think of a dozen other things at the same time, because the movements of my hands cutting slices of bread, taking milk and butter out of the refrigerator, etc, are not something I have to actively decide to do at each instant; they are responses to the information I receive through my retina about where all those things are in space, together with information stored in my memory and the reasons I have for eating this specific type of food right now (my preferences and beliefs about nutrition). Those reasons do not have to be selected each time I perform this action, they are standing reasons that I can act on each time I prepare my breakfast *ceteris paribus*. Note, however, that unlike the patient suffering from UB, a "normal" agent who acts spontaneously will not act on the stimuli she receives from the environment unless she has a previously acknowledged reason for doing that (e.g. a reason for hammering that nail to the wall).

[88] Churchland, P. (2013), p.154.

While we are responsive to perceptive signals, we are also responsive to our previously formed reasons, and among all the reasons we might have in favor of a certain action, only some were selected as our effective reasons for acting at this moment. That selection was made by the agent, through her power to form intentions to act. Even in spontaneous actions such as these, the exercising of this power is what renders us the authors of what we do, not a passive pawn in the game, or a victim of our "temper" or of the craving for a certain drug.

The next level of agency regards what I call ***actions-on-the-spot***. These are actions that result from fast decisions: there is still no conscious process of deliberation involved but, unlike in the case of spontaneous actions, there is some uncertainty about the outcome and so the intervention of the agent involves deciding which action to perform. One famous example of this type of actions are the actions performed by subjects in Libet-type experiments. In this family of experiments, conceived originally by Benjamin Libet and colleagues[89] but replicated by numerous other scientists[90], subjects are asked to flex their wrist or to press one or more buttons, whenever they feel like it. Their decision might regard only the timing of the action (in the case of wrist flexing or pressing just one button) or both the timing and the action itself (in the case of choosing which of two buttons to press). In both cases, however, it is clear that the agent is put before different options (for example, at each instant she has to decide whether to press a button or not) and she is asked to decide on the spot. Subjects in these types of experiments are asked not to plan their actions in advance, and so what they will eventually do depends on a fast decision that is not preceded by a process of deliberation.

This is analogous to what happens when we make immediate decisions in our everyday life, such as deciding which pair of socks to put on in the morning or which pack of Pampers diapers number 6 to pick at the supermarket. Admitting there is no reason to prefer one pack to another, nor to use the blue woolen socks rather than the green ones on a cold day in which I will be wearing boots, the decision is preceded by no deliberation,

[89] Libet, B. Gleason, C.A., Wright E.W., Pearl D.K. (1983), mentioned in section 1.2.

[90] Most famously Soon C.S., Brass M., Heinze H.-J., Haynes J.-D. (2008) and Fried I., Mukamel R., Kreiman G. (2011), mentioned in section 1.2.

even though it is made and authored by the agent, who is influenced by previous habits and tendencies. Also in the case of actions-on-the-spot, agency is common to human and many non-human animals. Superior creatures like dogs, mice, sharks and snakes are obviously capable of making fast decisions when they are faced with alternatives. They can choose this path or that, they can go fetch some food or sit a little longer in the sun, they can chase a difficult prey or procrastinate. Like Helen Steward has pointed out:

> "If one watches a large farm animal, such as a cow or a sheep, engaged in its normal activities, it is almost impossible, I suggest, for a normal and unprejudiced human being to avoid looking upon it as an agent. One supposes, that is, that though nature may have prescribed for it a number of essential activities (grazing, mastication, sex, drinking of water, etc.) from which it is certainly not free to forbear, *it* nevertheless determines the details of how, when, and where exactly these activities are to be carried out."[91]

Last but not the least, ***deliberative actions*** are at the top of the pyramid. They regard those processes of choosing (and acting upon one's choice) on the basis of reasons consciously weighed by an agent. When we think about which road to take to avoid traffic or whether to accept a new job far away from home, we take (more or less) time to consider the different options we have, the reasons for and against each option, and make a commitment for one of them based on these reasons. The main difference between the aforementioned actions-on-the-spot and deliberative actions is that in the former case reasons influence the choice the agent makes without she/it consciously considering them in advance, whereas in the latter case the reflective mind of the agent mediates the decision by pondering upon those reasons explicitly. This of course can happen very fast, maybe in a fraction of a second.

This last degree of agency is human exclusive, so far as we can tell. It requires abstract reflection, something for which a high level of brain complexity is required. However, one must not forget that all the elements that distinguish the "lower" forms of agency from simpler behaviors are present in deliberative actions as well, namely, the ability to form

[91] Steward, H. (2014), p.75.

intentions to act, reasons-responsiveness, and the capacity for decision-making. It is important to keep this in mind so that one does not make the mistake of inferring that the criterion for deliberative agency is simply reasoning. Instead, if we first analyze the requirements of agency as such, from its simplest to its most complex variations, it will become clear that agency is not possible without an irreducible entity endowed with downward causal powers.

2.3. The contribution of the agent

What happens when someone acts? This is the question David Velleman asked in his homonymous article of 1992. In that influential paper, Velleman presented an important objection to the davidsonian model of action production, according to which, under normal circumstances, beliefs and desires jointly cause the intention to perform a certain action. Velleman argues:

> "In this story, reasons cause an intention, and an intention causes bodily movements, but nobody – that is, no person – *does* anything. Psychological and physiological events take place inside a person, but the person serves merely as the arena for these events: he takes no active part."[92]

I follow Velleman's lead here, even though his analysis is restricted to what he calls "full-blooded actions" (the equivalent of deliberative actions), whereas I believe it could be extended to every form of agency. When an agent acts, her role can be described as: 1) forming an intention to act for certain reasons and 2) producing a bodily movement according to that intention. It is not the reasons that produce the intention *per se*, nor the intention that produces the movement. If that were the correct story, then we would be unable to justify why, under some circumstances (which might be abnormal, like the pathologies we analyzed in the previous section, as well as quite regular, as in cases of absent-minded or emotionally driven behavior), the agent can fail to participate in the

[92] Velleman, D. (1992), p.461.

behavior that she is supposed to be the author of. After all, in all these cases reasons and intentions would be present just the same.

Like I said in the previous section, the difference between the behavior of addicts and non-addicts, for example, is not that the former is not driven by reasons. Addicted behavior *is* motivated by certain desires (for the drug) and beliefs (that shooting up a certain dose in a certain manner will provide the desired effect). The problem is that in such cases there is no intermediate agential intervention between the reasons and the intention, nor between the intention and the behavior. The drug addict goes about his business in autopilot mode, like a kleptomaniac or a manipulated person, and this is why his behavior cannot yet be considered to be an action.

Velleman's proposed way out of this problem is event-causal and reductionistic.

> "My objection to the standard story is not that it mentions mental occurrences in the agent instead of the agent himself; my objection is that the occurrences it mentions in the agent are no more than occurrences in him, because their involvement in an action does not add up to the agent's being involved."[93]

Velleman's view is that agency cannot be reduced neither to reasons nor to intentions, because the agent intervenes *between* these elements. However, it would be question-begging to assume that the agential intervention cannot consist in the occurrence of certain mental states. Since a naturalistic explanation, according to his view, must involve states and events[94], what he suggests one should do, before giving in to a substance-causal or dualistic alternative, is to look for an extra element that might play the causal role of the agent in cases of full-blooded action. He believes to have found such an element in a person's desire to act for reasons:

> "The agent, in his capacity as agent, is that party who is always behind, and never in front of, the lens of critical reflection, no matter where in the hierarchy of motives it turns.

[93] *Idem*, p.463.

[94] He is not clear about what he means by "events" or "occurrences", but we can assume that he takes an event to be "a particular's having a property at a time, or standing in a relation to another particular at a time" [Clarke, R. (2003), p.155].

> What mental event or state might play this role of always directing but never undergoing such scrutiny? It can only be a motive that drives practical thought itself. (…) What animates practical thought is a concern for acting in accordance with reasons. And I suggest that we think of this concern as embodied in a desire that drives practical thought."[95]

Can we follow Velleman in this reduction of the causal role of the agent to the desire to act for reasons, the desire to do what makes sense and is intelligible to her? I believe there are two main problems with this proposal. I have already mentioned the first problem, which is that Velleman's account concerns only "full-blooded" actions, human actions *par excellence*. This is problematic because the element he identified as crucial cannot be adopted as definitory of agential intervention *per se*, from the level of spontaneous actions, through actions-on-the-spot, to deliberative actions. This is a drawback of his account because the degree of agential control with which I perform a daily routine action such as brushing my teeth is both very similar to the control I have when I deliberatively choose what to wear at an important meeting, and radically different from the degree of control a kleptomaniac has when she steals a certain object from her best friend. These similarities and differences are something that has to be accounted for.

Also related to this first problem is the fact that Velleman chose to restrict his account to deliberative actions because he takes for granted that in many other cases (that "lack what distinguishes human action from other animal behaviour"[96]), people might perform an action without taking active part in it. This strikes me as incoherent for two reasons: On the one hand, because an action in which the agent is not active sounds like an oxymoron (even if philosophers can have very different positions concerning what the active participation of the agent amounts to). On the other, because if there is no difference between certain human actions and other animal behaviors, then either the putative human actions are actually behaviors or the so called animal behaviors are actually actions. Should the fact that a certain undertaking is performed by a human

[95] *Idem*, pp.477-478.
[96] *Idem*, p. 462.

rather than a non-human animal suffice to define it as an action? I do not think so, and I hope to have made my case for this point in the previous section.

The second problem with Velleman's reductive account of the agential role as the desire to act for reasons is that it is too narrow. An action can be authored by the agent regardless of how the agent can herself explain it. It is entirely coherent to conceive the experience of doing things that we do not understand but which we fully endorse as actions that we performed intentionally: many of us (if not all) have had this experience some time or another. Also, we often act akratically, acting from reasons that we do not consider to be our best, and this does not prevent us from being accountable, nor others from giving us credit or blame for our acts – which reveals how much we are considered to have actively contributed to the action's coming to be.

Nomy Arpaly and Timothy Schroeder[97] have made a similar objection to Velleman's account with the help of Mark Twain's famous character Huckleberry Finn. In what they qualify as an act of moral inverse *akrasia* (a situation in which the agent acts against what she considers best but, due to her lack of judgment, the akratic course of action is more praiseworthy than the alternative one), Huckleberry fails to return Jim, a runaway slave he became friends with, to Miss Watson, his lawful owner. Because of the moral principles he was taught in rural Missouri, Huck believes returning Jim is the right thing to do and blames himself as a weak and bad boy for not being capable of acting morally, according to his convictions:

> "Conscience says to me: 'What had poor Miss Watson done to you, that you could see her nigger go off right under your eyes and never say one single word? What did that poor old woman do to you, that you could treat her so mean?...' I got to feeling so mean and so miserable I most wished I was dead"[98]

Even though weakness of will and moral weakness are two distinct phenomena[99], they coincide in this case, given that Huckleberry is in fact failing to do something he believes

[97] Arpaly, N., Schroeder, T. (1997). In the use of this example, the authors followed the lead of Jonathan Bennett's (1974).

[98] Twain, M. (1885), quotations taken from the online edition at e-booksdirectory.com.

[99] Cf. Zilhão, A. (2005).

to be both the right and the best thing to do. He wants to denounce Jim and he plans on doing it:

> "My conscience got to stirring me up hotter than ever, until at last I says to it: 'Let up on me—it ain't too late, yet—I'll paddle ashore at first light, and tell.' I felt easy, and happy, and light as a feather, right off. All my troubles was gone."

However, Jim's trust in him (his words of goodbye: "Dah you goes, de ole true Huck; de on'y white genlman dat ever kep' his promise to ole Jim") make his emotions win over his rational decision. When the chance comes to tell on the runaway slave, he just cannot find the strength to do it:

> "I didn't answer up prompt. I tried to, but the words wouldn't come. I tried, for a second or two, to brace up and out with it, but I warn't man enough—hadn't the spunk of a rabbit. I see I was weakening; so I just give up trying."

What Arpaly and Schroeder argue is that Velleman's account fails to justify the common intuition that Huckleberry's action is praiseworthy, despite its akratic nature. In fact, according to Velleman, an agent is not accountable for acts contrary to the desire to act rationally; Huck's motivation is not grounded on a desire to act for reasons, it is driven by a gut feeling of sympathy for Jim, hence the boy deserves no praise for what he has done. Since I am not concerned with moral judgement, I will not pursue this argument. My point instead is that Velleman's view would prevent us from even considering Huck's act as a proper action, and this seems utterly unreasonable to me.

If Huckleberry is able to find many more reasons in favor of the action of returning Jim than in favor of the action of protecting him – which is, in fact, an alternative in favor of which he can find *no* reason at all – then, on Velleman's account, when he fails to act according to his reasons, he quits being an agent. There is no agential intervention in the causal sequence that leads to his bodily movements; they constitute a mere animal behavior. This is such a counterintuitive conclusion that we just have to give up the view that would force us to accept it. When in Huckleberry's difficult deliberation, sympathy ends up weighing more than morality, there is a clear intervention of his self. His desire

to act for reasons (a desire which he did have and intended to act on) fails to move him, because he is not passive, because he takes sides.

But is the misjudgment of this case as a sub-actional behavior a problem related to the specific attitude that Velleman chose as being functionally identical to the agent (the desire to act for reasons), or is it a more general drawback capable of affecting any similar type of account?

I believe that no reductionist account of the agent can appropriately respond to the problem that Velleman points out. He claims:

> "What makes us agents rather than mere subjects of behaviour – in our conception of ourselves, at least, if not in reality – is our perceived capacity to interpose ourselves into the course of events in such a way that the behavioural outcome is traceable directly to us."[100]

I agree entirely. But I suspect that, no matter what psychological states and events a reductionist might elect as the core elements that can "speak for the agent"[101], there will always be an available counterexample of an action that lacks that element in its etiology but which we are willing to count as an action nonetheless. What is it that grants us the "capacity to interpose ourselves into the course of events in such a way that the behavioural outcome is traceable directly to us" in *all* the events that we call actions? What is undoubtedly present in all actions is not a *state* but rather an *ability*: it is the agent's power to form an intention to act and, through that intention, to be the cause of the action – that is, her Will.

The will is a tricky concept in the philosophy of action. It has recently been associated with volitionism, a theory according to which what defines an action as such is that it is either identical or it begins with a basic mental action called a volition (which is the agent's willing or trying to move her body in a certain way). What motivates volitionists is the idea that it should not be an external element such as the action's causal history, to determine whether it is in fact an action or just an event. What characterizes basic actions should be

[100] *Idem*, pp.465-66.

[101] Bratman, M. (2007), p.4.

a noncausal element, such as an "actish phenomenal quality"[102] or some sort of intrinsic intentionality[103]. However, this view faces many problems, most importantly the idea that the agent (as well as her reasons) acts without causing the action. How does the action come about, in a physical world? As it has become clear so far, I do endorse a causalist view according to which the agent's reasons (as well as her intention) are among the causes of her action. So I will define the will in a way that is independent from volitionist theories.

In my view, the agent's will is her power to make decisions or to form intentions to act. This power is not a human exclusive ability. Like I have said before, spontaneous actions and actions-on-the-spot can be performed by any animal whose mental capacities include the possession of beliefs and desires, the formation of intentions and, sometimes, the making of decisions. I contend that most mammals, for instance, can be said to be able to act of their own will.

This may seem surprising to the philosophical community, a large portion of which tends to attribute propositional attitudes to creatures with language only[104]. However, once we engage in common interactions with animals (say dogs, cats, horses), we interpret their behavior on the basis of the assumption that they do what they do because of what they want, believe and intend. The general assumption in experimental neuroscience, for instance, is also that the evolutionary continuity in biological systems allows us to infer that animals share with us the ability to act for reasons despite their lack of language. For example, neuroscience has been studying the neural basis of decision-making for over a decade and the main trend is to assume that "the path from simple decisions to complex ones may be more straightforward than it appears"[105], since the models that have been developed in experiments in non-human primates as well as in rodents can apply successfully to humans[106].

[102] Ginet, C. (1990), p.13.

[103] McCann, H. (1998), p.163.

[104] Cf. Davidson, D. (1982).

[105] Gold, J.I., Shadlen, M.N. (2007), p.562.

[106] I am referring, for example, to the integration-to-bound model according to which the brain accumulates inputs (which may come from the environment, in evidence-based decisions, as well

However, it is important to note that to say that an animal or human agent acts of her own will is different from saying that she acts of her own *free* will. In fact, the question we must ask now is if all actions are equal in what concerns the agent's degree of control and freedom. This is what I will turn to in the next section.

2.4. Free action and free will

We have seen so far that in order for an action to be such, it has to be intentional. This implies that it must be the product of the agent's willful act of forming an intention to act in a certain way. This means that every action is a free action, in a certain sense: it is an event that is brought about by an agent in such a way that it was not made inevitable by any extrinsic cause (it was not coerced), nor by an intrinsic force (it was not compulsory); instead, it was appropriately caused by the agent's intention to act for certain reasons (beliefs and desires). The agent did what she did because she so decided or simply intended.

This is not the only way in which one can conceive of freedom, though. In the classical debate on the compatibility between determinism and free will, this sense of free action is not what is at the center of the controversy. An agent may perfectly well have the ability and the possibility to do what she wants, but the question incompatibilist philosophers ask is: could she have wanted otherwise? In other words: is her *will* free?

The difference between both these meanings of freedom became apparent in the famous seventeenth century debate between Thomas Hobbes and Bishop John Bramhall, in which the former argued:

---

as be caused by some inner source of variability, in spontaneous decisions) voting for or against an action, but only commits to a definite decision once a certain threshold is crossed. This model, which is practically consensual in what concerns animal evidence-based decisions [Roitman, J.D., Shadlen, M.N. (2002), Hanes, D.P., Schall, J.D. (1996), Krajbich, I. et al. (2010)], has recently been argued to provide an adequate explanation for the "readiness potential" (a typical neural pattern observed in a EEG) which precedes actions-on-the-spot in the abovementioned Libet experiments performed in humans [Murakami, M. et al. (2014), Rigato, J. et al. (2015)].

> "Liberty is the absence of all the impediments to action that are not contained in the nature and intrinsical quality of the agent. (…) I conceive that nothing takes beginning from itself, but from the action of some other immediate agent without itself. (…) So that whereas it is out of controversy that of voluntary actions the will is the necessary cause, and by this which is said the will is also caused by other things whereof it disposes not, it follows that voluntary actions have all of them necessary causes and therefore are necessitated."[107]

Determinism is the contemporary term for what Hobbes called "necessity": it is the causal nature of a world in which given a full description of all its elements and laws at t_1, only one possible state can follow at t_2[108]. In other words, a world in which given a certain physical cause, the physical effect becomes inevitable. It is clear, then, how Hobbes' freedom of action is independent from the truth of determinism. According to his *compatibilist* view, what is crucial is that the agent's intention to act is caused by her will, which is identified with her most effective desire. It becomes irrelevant whether her will is necessitated or not by previous causes. But Bramhall questioned this view: even if the agent may act as she wills, "if the will has no power over itself, the agent is no more than a staff on a man's hand"[109]:

> "Whosoever have power of election have true liberty, for the proper act of liberty is election. A spontaneity may consist with determination to one, as we see in children, fools, madmen, brute beasts, whose fancies are determined to those things which they act spontaneously, as the bees make honey, the spiders webs. But none of these have a liberty of election, which is an act of judgment and understanding, and cannot possibly consist with a determination to one."[110]

---

[107] Hobbes, T. (1654), *Of Liberty and Necessity*, in Chappell, V. (1999), p.38.

[108] Robert Bishop (2011) defines physical determinism on the basis of four conditions: differential dynamics, unique evolution, value determinateness and absolute prediction. I believe that the concept of Unique Evolution [although defined by Bishop in the language of physical science: "A model is such that a given state is always followed by the same state transitions" (p.85)] is the one that can capture best the feature of determinism that underlies its tension with free will: the existence of only one possible future, given the past and laws of nature.

[109] Bramhall, J. (1655), cit. in Chappell, V. (1999), p.44.

[110] From Bramhall's original "Discourse of liberty and necessity" (written in 1645, but unpublished until 1676), in Chappell, V. (1999), p.2.

Bramhall's ideas were actually a reflection of his scholastic inheritance (contrasting with Hobbes' modern materialism) and they are still present today, under different versions, in the *libertarian* theories of free will. One of the most famous ones is Robert Kane's, according to whom free will is "the power of agents to be the ultimate creators (or originators) and sustainers of their own ends or purposes"[111].

Under a view such as Hobbes', that reduces the agent's will to her "desires and inclinations" (or to her most effective one), the possession of these reasons is sufficient to cause the production of an intention to act accordingly. The intention is an intermediate mental state, caused by other mental states (beliefs and desires), leading to the action. The problem is that such an account faces the objection we developed in the previous section: it reduces the agent to the locus where the whole causal chain that leads from reasons to action takes place, and nothing more. Since the will is a state and not an ability, the agent's role is passive, for what brings about the action are the beliefs and desires that she *has*, not something that she actually *does* – and this seems to be an unsteady basis on which to ground the distinction between actions and behaviors.

It is in order to avoid this drawback that many contemporary philosophers of action[112] as well as myself endorse the view according to which the active element of forming an intention is crucial for an action to be considered such. As I have already explained, the will is the name given to this ability by which the agent intervenes in the causal chain by committing herself to a plan and hence actively contributing to the outcome.

This type of account is noncommittal with regard to the question of determinism, for what is crucial is the active forming of an intention, not that that intention-forming act can escape causal determinism. And since the agent's will does not have to be indeterministic in nature for an intentional behavior (an action) to be clearly distinguished from an unintentional one, there seems to be no reason to think that action cannot take place in a deterministic world. The agent can act even if her will could not have been different

[111] Kane, R. (1998), p.4.

[112] Cf. Bratman, M. (1987) and Mele, A. (1992).

from what it is. So we can conclude that actions are always free (in a compatibilitist sense), but need not be free willed (in a libertarian sense).

However, even if the structure of action as such does not imply that the agent's will be free, the degree of control the agent possesses over her action increases when it is in fact free, for she becomes able to settle not only what she will do but also what she wants to do. John Martin Fisher established an important distinction between these two types of control, which he calls respectively "guidance control" (a reasons-responsive mechanism that is the agent's own) and "regulative control" (which involves the ability to choose and act differently under the exact same circumstances).

> "[S]uppose you are at the controls of an airplane, a glider, and you are guiding the plane to the west. (…) You consider whether to steer the plane to the east, but you decide to keep guiding it to the west, in part because the scenery is nicer in the west. Unknown to you, the wind currents in the area are such that the plane would continue to go to the west, in just the way it actually goes, even if you had tried to steer it in some other direction. (…) In this example, you steer the plane to the west in the "normal" way. It is not just that you cause it to go to the west (which you would equally have done had you steered the plane in the same way as a result of a sneeze or an epileptic seizure). Rather, you guide the plane in a distinctive way — you exhibit a signature sort of control, which I shall call "guidance control." Here you exhibit guidance control of the plane's movements, but you do not possess regulative control *over* the plane's movements."[113]

Regulative control is the type of control that Bramhall aimed at, one that gives the agent that "liberty of election" by which she can be in charge of her own will notwithstanding the past and present circumstances, her character and reasons. An action that proceeds from this level of control is different from an action over which the agent possesses only guidance control because in the former case there are open alternative futures and the agent has the power to determine which becomes actual, whereas in the latter case there are not. However, the possibility of someone acting with a libertarian type of control is

[113] Fischer, J.M. (2006), p.8.

still under heavy controversy because of its questionable coherence and (according to many) scientific implausibility.

These are topics I will assess only further ahead in this dissertation. For now, I am concerned with other important implications of the capacity of forming an intention to act *per se*, which might give us a surprising new perspective on the compatibilism/ incompatibilism debate.

2.5. A non-aggregational agent

If the active role of the agent in raising her arm marks out a contrast with what happens to her in cases of unwilled behaviors, then the existence of an intention-forming entity responsible for the mental act of forming an intention to act must be assumed. And here is where agent-causalism, a theory that has been met with much undeserved suspicion, enters the picture.

If the agent were nothing more than the mereological sum of her mental states and events and their neural correlates (a "bundle or collection of different perceptions"[114], quoting Hume) and her intending could be reduced to her intention being brought about by some of those, then there would be nothing, besides them, which might influence the behavioral outcome. On such an account, the agent would in fact be just a name given to a collective entity, which would be constituted and controlled by its parts (her reasons and other intentional states), without in turn being able to control them, whether they were necessitated or not by past events.

A “humean” agent is a composite entity with structural properties[115], just like rocks or plants are, and thus her bringing it about that she will do A instead of B is actually the

[114] Hume, D. (1738), *A Treatise of Human Nature,* 1.4.6.4 (the Norton edition, the last number indicating the paragraph).

[115] Cf. O’Connor’s definition of structural properties: "A property, S, is structural if and only if proper parts of particulars having S have some property or properties not identical with S, and this state of affairs is constitutive of the state of affairs of the particular's having S." [O’Connor, T. (2000), p.109].

result of each of her parts' causing a certain complex collection of events at the mental as well as the neural level. Thus for an action to be brought about by the agent, as opposed to her being passive relative to the occurrences taking place within her, the agent as such must be the cause of the action, by willing. This means that at the psychological level as well as at the neurophysiological one, the agent whose will determines the outcome must be something more than some of, or all, her mental states.

Let us go back to some of the examples we used in previous sections in order to understand this better. When one compares the compelled behavior of a kleptomaniac, a drug addict or a manipulated person, on the one hand, with the willful acts of a person under normal conditions, on the other, one realizes that the main difference between them lies in the degree of control that connects the agent with her action. What the first three cases have in common is that they describe people that are not in charge of their behavior, even if it flows out from their inner mental states. On the contrary, the regular agent is someone whose autonomous will can supersede the blindness of event causality which would otherwise make her an automaton. Both compatibilists and incompatibilists agree that self-determined action entails physical and moral responsibility insofar as it involves control, and control is another word for authorship. If the agent were not the author of the action, her behavior would be something that *happens* to her and not something she *does*. The agent's power lies in the fact that she *herself* (and not the psychological and neural events that *happen within* her) is the action's author.

Therefore, any view that recognizes the importance of the intention-forming act requires the postulation of a "non-aggregational"[116] self, who is endowed with the ability to commit to a plan based on (but not necessitated by) the reasons it has, and whose causal power to so intend is not merely derivative. Such a view is clearly an agent-causal view: one that considers the agent as a fundamental cause of the action, a necessary cause that cannot be described in event-causal terms[117].

---

state of affairs of the particular's having S.

[116] I am borrowing this term from Clarke, R. (n.d.).

[117] Whether the agent-causalist will endorse an account of the agent that defines the human person as an animal, a brain, a soul or something else, is something that I believe can be left

Many event-causalists counter this view by arguing that there is a way to avoid passivity without having to postulate the agent's ability to intervene via some sort of emergent and downwardly effective causal power. Like Velleman's "desire to act for reasons" that I have already presented here, other solutions proposed by Frankfurt[118], Watson[119] or Bratman[120] purport to show that if the agent (as a complex psychophysical system) functionally *identifies* with some of her states and these states play the self-determining causal role in bringing about the action, then, thanks to that identification, it is *as though* the action was directly caused by the agent as such. What is needed, these authors say, is that the agent's effective reasons are recognized by the agent's evaluative system as valid and thus fully endorsed.

I do not deem this suggestion to be capable of adequately distinguishing a free agent from someone who is a victim of addiction, for instance. The identification of the agent with the desire that caused her to act could have happened merely by chance! Imagine the case of a kleptomaniac who defends the righteousness of theft and conscientiously wishes to steal: when he falls prey to his compulsive behavior, he is no more in control of it than some other kleptomaniac who does not want to steal. It does not matter if he might agree or identify with his action: what does matter is what eventually caused it[121].

To conclude, agents whose free actions are internally caused by their psychophysical states in the same manner as their compulsory or unconscious actions are, cannot be considered agents at all. This implies a non-aggregational account of their authoring self, which must be irreducible to any mental states and events that might have the power to

unsettled for now. Unlike what is sometimes thought, an agent-causal view does not need to presuppose some spiritual being in order to have theoretical coherence, it only needs an irreducible substance (which can be something as concrete as a living and thinking brain) whose power to act amounts to more than the sum of the powers of its parts. Timothy O'Connor (2000), Randolph Clarke (2003), Jonathan Lowe (2008) and, more recently, Helen Steward (2014) are all examples of authors who tried to put forward accounts of agent-causation that are compatible with a naturalistic stance.

[118] Frankfurt, H. (1971).

[119] Watson, G. (1975).

[120] Bratman, M. (2000, 2005, 2007).

[121] This case is very similar to that of Frankfurt's willing addict, that I have already mentioned here (Frankfurt, H. (1971).

bring about their bodily movements in cases of sub-actional behavior, while at the same time being present and active in all those cases that we are willing to consider proper actions.

2.6. Agent-causal libertarianism

There is a peculiarity in my analysis so far that might have caught the attention of some of my readers: I am not following the standard approach according to which compatibilist accounts of agency always reduce the agent's causing the action to events involving her causing it, while libertarian accounts might either do that or not. In the usual taxonomy, in fact, only libertarians are given the option between event-causal incompatibilist accounts or the less popular agent-causal accounts, which have been heavily criticized as ontologically obscure and scientifically implausible.

As I have pointed out, any reductionist view about agency falls short of endowing the agent with enough authorship over what she is doing. If an account of action does not leave room for the agent's intervention, it will be unable to tell apart sub-actional behavior from proper actions and to justify the intuitive belief according to which these are different sorts of things. This entails that agent-causalism, i.e. the view that includes the agent as a causal element apart from her mental states and events (or their neural correlates), can and should be defended by both compatibilists and incompatibilists[122].

Libertarian agent-causalists, however, have an extra argument against event-causal versions of their contention that free action implies an indeterministic type of free will. In order to understand this argument, let me first present briefly how their position is framed in the contemporary debate about free will.

According to any *incompatibilist* view, the agent can be said to act freely only if she has the ability to make different choices under the exact same circumstances, given the past and laws of nature. This means that, if the film of the universe were to be replayed all

[122] Helen Steward (2014) is one of the few who has defended such an account. Other authors have discussed it, such as Randolph Clarke (2003, pp.163-3) and Christopher Franklin (forthcoming c).

over again, from the Big Bang up until now, and every single detail were to happen exactly the same way, the agent might still act otherwise here and now and thus give rise to alternative futures.

Libertarianism is the name given to all forms of non-skeptical incompatibilism and event-causalism is its most popular version. It relies on the idea that all causes are events and on a causal theory of action according to which "an event is taken to be an action in virtue of being caused in a certain way by mental events of certain sorts"[123]. It differs from compatibilism only in the requirement that the causal relation between those mental events and the agent's action be undetermined, i.e., that there be genuine open alternatives up to the moment when an intention is formed. As in any other cases of indeterministic causation, the effect is underdetermined by its causes and so it is not empirically necessary but only probable. In the specific case of free decision, all the reasons the agent acted on, and were thus effective in producing that particular action, did cause it; they made it possible, but not inexorable.

In contrast, according to agent-causal libertarianism, there is an extra crucial element that must enter the etiology of action: the agent, who is the one who "tips the balance"[124] in cases of torn decisions and the one up to whom the ultimate choice is, even in cases of unbalanced options. One of the main objections that agent-causal libertarians (as well as some nihilists) make to their event-causal counterpart is the so called "disappearing agent argument"[125] (which is similar to Velleman's objection to the davidsonian picture of action production, but with a further element regarding chance). The argument can be stated as follows:

[123] Clarke, R. (2003), p.25.

[124] This image is used by Carl Ginet, criticizing Clarke's view on the collaborative causation of free action by reasons and the agent as a substance, in Kane, R., ed. (2002), p. 398.

[125] Cf. Pereboom, D. (2001, 2004, 2007, 2012); Griffith, M. (2010). The "No-Choice argument" by Peter Van Inwagen argues in a similar manner for the agent's insufficient control over his own decision: "If an agent's act was caused but not determined by his prior inner state, and if nothing besides that inner state was causally relevant to the agent's act, then that agent had no choice about whether that inner state was followed by that act" [van Inwagen, P. (1983), p.149, cit. in Clarke, R. (2003), p.98].

> DISAPPEARING AGENT - *If the decision remains genuinely undetermined up to the very last moment, and if the agent can be reduced to her psychophysical states and events, which ultimately bring about that decision, then she has an insufficient control over which decision is eventually made.*

This is a very strong consequence of both the first presupposition of any incompatibilist account (that the decision is indeterministically caused) and event-causalism (the idea that everything that happens in the world is caused by occurrences rather than by objects). As Derk Pereboom framed it:

> "With the causal role of the antecedent events already given, whether the decision occurs is not settled by any causal factor involving the agent. In fact, given the causal role of all causally relevant antecedent events, nothing settles whether the decision occurs."[126]

The final decision happens merely by chance, and the agent 'disappears' from the causal etiology of action. This is hardly something a libertarian should feel comfortable with. However, according to agent-causalists and other critics of event-causalism, the disappearing agent problem arises only if the libertarian should assume a reductionist account of the agent[127] together with an event-causal view. As Timothy O'Connor, one of the few contemporary agent-causalists, illuminatingly put it:

> "Even though the [event-causal libertarian] account allows for the real possibility of different courses of action, any of which would be 'controlled' by the agent in the minimal sense of being an 'outflowing' of the agent, it's not 'up to the agent,' something he 'has a choice about,' just which potential cause will be efficacious in any given instance and so which action will actually occur. It is, rather, a matter of its falling under a statistical or quasi-statistical tendency that governs the general pattern of behavior in types of

[126] Pereboom, D. (2014b), p.61.

[127] Even though they usually do, it is important to note that neither compatibilists nor event-causal libertarians *have to* commit to a reductionist account of the agent. One thing is the answer to this metaphysical question: "which kinds of things are causes?"; another is the answer to this question of basic ontology: "what fundamental things are there?"; another still is the answer to questions related to action theory: "What is an action and what sort of thing is an agent?". While the event-causal libertarian gives, of course, an event-causal answer to the first question, she could nevertheless consider the agent to be an irreducible substance. Just not a substance-cause.

> circumstance over time, and this probabilistic tendency clearly is not something the agent has any choice about". [128]

For an action to be freely brought about by the agent, she must have the capacity to settle which of the possible alternatives becomes actual, and this depends both on the metaphysics of causation that is true of our world (or at least of action production), and on the ontology that defines what an agent is. On the one hand, it must be the agent, as the bearer of certain properties, that causes an action to occur, as opposed to it being the event of her having, coming-to-have or ceasing-to-have a certain property or causal power at a certain time. On the other, this agent must be conceived in a non-humean way. In fact, under an aggregationist or humean ontology, the causing that is experienced by the agent as a whole is actually the result of many tiny causal processes and that, as we have seen above, would prevent us from adequately distinguishing actions from non-actional behaviors. A substance-causal metaphysics is not sufficient for agent-causation, since the irreducibility of the agent's causal powers to events involving her does not entail the irreducibility of herself as a substance-cause.

To sum up, a libertarian theory that wishes to enhance the agent's control over her action must make two major modifications to its underlying metaphysics:

1) It will have to abandon the idea that all causes are states or events, for an alternative view that recognizes that causes of (at least[129]) actions are substances. This will allow it to accept that the agent herself, as an irreducible substance, is the ultimate cause of her actions.

2) For that second clause to be true, the agent-causalist will also have to substitute robust accounts of the self (emergentist or dualist), for the bundle approaches that reduce the agent to the sum of her parts.

[128] O'Connor, T. (2000), p.29.

[129] E.J. Lowe, for example, endorses the view according to which *all* causes are substances: "Events, in my view, may be said to be causes at best only in a loose and derivative sense, as a convenient *façon de parler*." [Lowe, E. J. (2008), p.5].

These are the same two modifications that I have contended *any* realist account of action (even a compatibilist one) must make. The second of these (the need for an irreducible self) appears to be in contrast with the reductionist research program that has allowed for the extremely detailed knowledge we have today about the mechanisms underlying behavior in biological systems. This knowledge is improving at a very fast pace and any philosophical account of action or free will must beware of assuming empirical commitments that may easily be refuted. Nevertheless, I am willing to embrace the challenge of understanding how agency might be possible in the physical world, as it is presented to us by the natural sciences.

As we have seen in the beginning of this chapter, scientific accounts too assume actions to be distinct from mere reflexes in that they manifest some sort of freedom from immediacy and presuppose a self that is in control. I believe what I have argued so far shows that these assumptions are incompatible with a reductionist account of action production. But is reductionism not an assumption as well?

My point is that, in order to avoid this apparent contradiction between a non-deflationist account of agency and the truth of reductionist physicalism, we must analyze better the assumptions upon which science stands and see how solid and indispensable they are. The problems of action, the mind-body relation, causation and the ontology of hierarchical levels in the organization of physical reality are all intertwined. As Helen Steward pointed out:

> "An answer to [the question how agency is possible] will require also an understanding of what could lead us to want to say that an organism rather than merely some part of one, or some process within one, has brought something about, and of how the causality thereby effected (the causality that is agency) relates to the causality involved in the sub-personal processes that make it possible."[130]

Ultimately, for both philosophers and scientists, it is worth asking these questions: How can a macro substance have downward causal effects over the undertakings of the parts

[130] Steward, H. (2014), p.11.

that constitute her? In the concrete case of human action, how can we say that the agent as such is the non-derivative cause of her bodily movements?

These are some of the questions I will address in the next two chapters.

3. IRREDUCIBILITY IN NATURE

> "Given the advent of quantum mechanics and these other scientific theories, there seems not a scintilla of evidence that there are emergent causal powers or laws in the sense in question… and there seems not a scintilla of evidence that there is downward causation from the psychological, biological and chemical levels."
>
> (Brian McLaughlin, 1992)

3.1. Definition and History of Emergence

Emergentism, in its various forms, is the view according to which there are features of reality that are irreducible to the lower-level basis from which they emerge, in the sense that they are more than just the result of the combination of the system's parts and their interactions. These features are paradoxically (or so it seems) both dependent on and autonomous from their emergence base, i.e. from the lower-level that brings them about.

This type of theory has been developed in very many areas, from different scientific branches to philosophy. In the latter field, its study stems from metaphysics and the philosophy of mind (in which the focus has been mainly on the relationship between mental/conscious entities and their physiological substrate), to philosophy of science in general (where the focus is on whether certain theories are reducible to others or not), to philosophy of physics (where emergence seems to be a good conceptual tool for explaining nonlinear phenomena) and philosophy of biology in particular (where the main interest is top-down causation in self-organizing systems). Needless to say, there is no uniformity in the way the concept of emergence is used and in the candidates that are accepted as good examples of emergent phenomena.

I will now present briefly the history of this concept in philosophy and frame the account I believe is useful and relevant for the discussion that concerns me in this dissertation.

"British emergentism"[131] was very popular in the second half of the nineteenth century and the first quarter of the twentieth century. Authors like Samuel Alexander[132], C. D. Broad[133] and Lloyd Morgan[134], among many others, endeavored to develop an account of phenomena like life or the mind that could be somehow intermediate between Mechanism and Vitalism. The former position seemed insufficient to explain the surprisingly novel properties that arise in nature at certain levels of complexity, whereas the latter was a dualist alternative which gave up on the goal of explaining how those properties could fit the natural order. The so called classical emergentists chose a third path instead, in which monism was rendered compatible with the irreducibility of special features like life or consciousness to the microphysical:

> "Put in abstract terms the emergent theory asserts that there are certain wholes, composed (say) of constituents A, B, and C in a relation R to each other; that all wholes composed of constituents of the same kind as A, B, and C in relations of the same kind as R have certain characteristic properties; that A, B, and C are capable of occurring in other kinds of complex where the relation is not of the same kind as R; and that the characteristic properties of the whole R(A, B, C) cannot, even in theory, be deduced from the most complete knowledge of the properties of A, B, and C in isolation or in other wholes which are not of the form R(A, B, C)."[135]

This means that the relations the primary constituents are involved in are what determines the systemic properties of the whole they will compose. Each different case of emergence manifests a "*unique* and *irreducible* law"[136], which can be learnt only empirically, not deductively, and which is what renders the systemic properties novel, despite their dependence on the substrate from which they arise. They are produced by

[131] Cf. McLaughlin, B.P. (1992).
[132] Alexander, S. (1920).
[133] Broad, C.D. (1925).
[134] Morgan, C.L. (1923).
[135] Broad, C. D. (1925), p.61.
[136] *Idem*, p.68.

the underlying elements and their arrangements, as they could not exist without them, but they are qualitatively new and unpredictable *a priori*. Whenever they appear, the general "principle of composition of causes"[137], according to which the effects are proportional to the causes, fails, which means that systemic properties are no longer resultant (additive).

Ingenious as this proposal might be, its classical versions were very problematic. When faced with the need to explain how this new type of relation worked, classical emergentists had no positive account to offer, and considered emergent entities as just a brute fact, a fundamental and irreducible expression of the laws of nature:

> "The higher quality emerges from the lower level of existence and has its roots therein, but it emerges therefrom, and it does not belong to that level, but constitutes its possessor a new order of existent with its special laws of behaviour. The existence of emergent qualities thus described is something to be noted, as some would say, under the compulsion of brute empirical fact, or, as I should prefer to say in less harsh terms, to be accepted with the 'natural piety' of the investigator. It admits no explanation."[138]

However, stating the unexplainability of certain facts does not illuminate the very problem that unexplainability represents and this has led many to consider that emergence was just another way of casting spooky phenomena out of the realm of natural science. Also, scientific improvements like the discovery of the DNA and the quantum mechanical explanation of chemical bonding dictated the fall of British emergentism, as it was too much based on the contingent lack of knowledge of nineteenth century science for it to be able to survive the incredible explanatory power of new reductive theories[139].

Recently, the concept of emergence has regained attention and the merits of its applicability to various fields have been defended with more sophisticated examples and arguments, both among scientists as well as philosophers of science, philosophers of mind

[137] Cf. John Stuart Mill (1868).
[138] Alexander, S. (1920), vol.II, pp.46-7.
[139] Cf. McLaughlin, B.P. (1992).

and metaphysicians[140]. The interest grew back due to various reasons, among which the developments in condensed matter physics which unveiled the impossibility of deducing the phenomena observed at the level of many-body physics by extrapolation from the properties of single particles (as in examples of spontaneous symmetry breaking). The implications of this problem for the myth of reductionism were famously explicated in a very influential article by Nobel laureate Philip Anderson entitled "More is different".

> "The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe. (…) Instead, at each level of complexity entirely new properties appear, and the understanding of the new behaviors requires research which I think is as fundamental in its nature as any other."[141]

Progresses keep being made in condensed matter physics, in the sense of exploring the philosophical consequences of some formal techniques aimed at explaining phenomena such as phase transitions, universality, and changes of scale[142]. But other fields in physical science have also contributed to the revival of emergentism, such as the physics of complex systems[143] and quantum mechanics (which has sometimes been considered almost a paradigmatic case of irreducibility because of the holistic nature of quantum entanglement[144]). Also in some areas of philosophy, advances were favorable to the reclamation of the emergence hypothesis, such as in philosophy of science (with the putative failure of Nagelian forms of reduction[145]), philosophy of mind (because of the persistent dissatisfaction of many with reductive solutions to the mind-body problem[146])

[140] Cf. the interdisciplinary research developed presently at the University of Durham, founded by the Templeton Foundation (Durham Emergence Project).

[141] Anderson, P.W. (1972), p.222.

[142] Cf. Batterman, R. (2002).

[143] Cf. Hooker, C. (2011).

[144] Cf. Hütterman, A. (2005).

[145] Ernest Nagel (1961) provided a formal model of the explanatory unification of the different sciences, according to which the laws of the reducing theory would deductively entail the laws of the reduced theory, but no examples of such inter-theoretic reduction were ever accepted universally.

[146] Cf. Nagel, T. (1974), Jackson, F. (1982, 1986), Levine, J. (1983), Chalmers, D. (1996).

and metaphysics (in which recent dispositional[147] and substance-causal theories of causation[148] have given new motivation to the search for a coherent and plausible theory of emergence).

For all these reasons, emergence is now considered a hot topic in philosophy and certainly not an outdated hypothesis in many scientific fields.

Before I move on to section 3.2. in which I explore putative cases of emergence in physics, it is useful to distinguish between epistemological and ontological types of emergence. The former type is uncontroversial and regards the relation between theories and the impossibility to deduce higher-level laws or properties from our knowledge of the lower-level elements and of the laws that govern their behavior and their relations. This might be only a practical impossibility that derives from the imperfections of our epistemic access to reality.

In turn, ontological emergence regards the relation between entities themselves (beyond the theories that describe them). It is an in-principle insufficiency of the lower-level domain (for example, the neurobiological states, events and laws) to provide a satisfactory explanation for the novel entities and features that appear at the upper-level domain (for example, consciousness). These entities emerge from the lower-level but are autonomous in the sense that their behavior is not reducible to the underlying properties and laws (i.e. it is not derivable from nor identical to them).

How can this in-principle irreducibility be established? We will get to that point in the next chapter, where I will attempt to show that we have independent reasons for considering that there are some phenomena in the world that can never be reduced to particle physics. However, regardless of those independent reasons, we must first put forward a more detailed account of what we mean by ontological emergence.

What is most important about the possibility of ontologically emergent entities (properties or substances) is that they carry with them new causal powers that are not derivative, i.e. that are basic (despite belonging to composite individuals). O'Connor and

[147] Cf. Mumford, S., Anjum, R.L. (2011), Jacobs, J.D., O'Connor, T. (2012).
[148] Cf. Lowe, E.J. (2008).

Wong have defined basic properties as "those properties whose instantiation does not even partly consist in the instantiation of distinct properties by the entity or its parts". And they add:

> "It is the thesis of emergentism that some basic properties are had by composite individuals."[149]

Unlike structural properties[150], the causal powers of which just amount to the collective entity's parts having certain properties and relating to each other in a certain way, emergent causal powers are not reducible to the causal powers of their emergence base. This means that the causal effects the upper-level entity can have are distinct from the effects its emergence base can have, even if we take into account all its parts and their relations.

For something to be correctly labeled as an emergent, it has to be both novel and real. The criterion most emergentists (myself included) assume to determine whether something is real (rather than just a collective entity that can be identified at a certain level of description but which has no ontological status of its own) is the so-called Alexander's dictum: *To be is to have causal powers*. Therefore, an emergent entity cannot be epiphenomenal – it has to be able to have causal effects in the world –, and it has to be novel – which means that its causal efficacy as a whole has to amount to more than the sum of the causings of its parts. So emergence entails upper-level causation: which can be same-level causation (e.g. mental-to-mental) or downward causation (e.g. mental to-neural).

Not all philosophers agree with this statement, though. John Searle, for instance, distinguishes between "emergent 1" features and "emergent 2" features: The former are not explainable only on the basis of the system's components and their arrangements, and rather require the causal interactions between the elements to be part of their explanation (that is why they are emergent 1). The latter, however, have the added merit of possessing causal powers that cannot be explained by the causal interaction of the

[149] O'Connor and Wong (2005), p.664.

[150] I am borrowing this term and this distinction from O'Connor and Wong (2005, p.663).

system's components, which Searle recognizes is a necessary condition for there to be downward causation. But alas, there are no cases of emergence 2, according to him. Even consciousness is an emergent 1 feature, says Searle, and thereby causally reducible to its neurological substrate.

Like I said in the beginning, this is one of the problems with the current literature on emergence: there is an enormous heterogeneity among concepts, with cases that some consider to be paradigmatic of emergence being rejected as clear examples of reducibility by others. Therefore, we must be clear from the start about what we believe emergence is and what instead are to be considered mere situations of structural complexity.

I will then conclude this section with my definition of ontological emergence (the type of emergence I am concerned about):

> *Ontological emergence is a relation between entities (substances or properties) belonging to different levels of organization of reality (different lengths and time scales), in which the lower-level structure gives rise to the higher-level entity without the novel causal properties manifested at the emergent level being reducible to the ones below.*

What does it mean to say that the emergent causal properties are irreducible?

> *For the causal properties of the emergent entity to be irreducible is for them to be both distinct from and causally unexplainable in terms of the properties of the constituent or subvenient*[151] *elements that originate them, nor in terms of the laws that govern their behavior and interactions*.

Several clarifications are in order. First, I refer to lower-level elements as "constituent or subvenient" because it is important that the entities that form the emergence base be

[151] Subvenient entities (intended here as *naturally* subvenient rather than *logically* subvenient – I will develop this distinction in section 3.6.2) are lower-level properties, events, substances or laws that are in a natural supervenience relation with upper-level entities. The natural supervenience of non-fundamental entities on the bottom-most metaphysical domain is the philosophical term for the asymmetric covariance relation by which all macro changes that those entities undergo depend on corresponding micro changes, but not vice-versa. Many authors believe that emergence implies the break of natural supervenience. I do not believe so and will argue against those who do in sections 3.6 to 3.8.

allowed to consist in mereological parts as well as not. The relationship between the brain and the conscious mind, for instance, is not a part-whole relation (thus failing to count as a constitution relation) but it is a very good candidate for an emergence relation nonetheless. Second, note that in the definition of ontological irreducibility there is an epistemological element that I failed to eliminate: explanation. In fact, it is not sufficient to claim that the causal powers of an emergent entity are distinct from those of its emergence base. The causal powers of a composite whole are very often distinct from those of its parts but if they derive from them, they cannot be considered emergent. The properties of common table salt, for example, are totally different from the properties of its components (sodium and chlorine) taken separately, but they are derivable from their bottom-level features plus the laws that govern their interactions. This means that the physical and chemical properties of the bottom-level structure are sufficient to explain the properties we can find in the compound. On the contrary, in cases of emergence, "it is not possible to *trace* the determinative chain that goes from the emergence basis to the emergent"[152].

The fact that we have to use the epistemological concept of explanation in the definition of ontological emergence is revealing of our limitations as subjects, of course. We cannot assess the degree to which a certain portion of reality (the emergence base) is a sufficient condition for the coming to be of another portion of reality (the emergent entity), independently from our epistemic access to both of them. Because of this, as we will see in next few sections, it is very hard to establish whether a certain entity is only epistemically or also ontologically emergent.

3.2. Emergence in Physics

A commendable tendency in the past years in philosophy of science has been to develop accounts of emergence that move away from armchair metaphysics and anchor philosophical analyzes in scientific theory and practice. Two important examples are

[152] Sartanaer, O. (2005), p.5.

Robert Bishop, a physicist and philosopher who has been working in questions of Emergence and Complex systems for many years, and Robert Batterman, a leading figure in Philosophy of Physics.

Alone[153] or together with Harald Atmanspacher (theoretical physicist)[154], and more recently with Michael Silberstein[155], Bishop developed an account of what he calls "contextual emergence", which is a relation between different levels of description (in its epistemological form) or between domains of reality (in its ontological form), whereby the description, properties or behaviors[156] of the lower domain provide some necessary but no sufficient conditions for the novelty existing at the upper-level. The remaining sufficient conditions must be provided by the context, which includes the emergent states and observables' stability conditions (that guarantee their existence and persistence), which are not given by lower-level descriptions. Bishop and his co-authors use several examples as evidence for the ubiquity of contextual emergence, from the domain of quantum chemistry to that of human society. All of them have to do with scale transformations: how the laws of microphysics give rise to the laws and properties of the macro world.

It is common knowledge that the **transition from quantum to classical mechanics** is a mysterious one. Mathematically, these two realms are separated by singular limits (mathematical expansions in which some quantities are assumed to tend either to zero or to infinity), which means that the transition between the formalism that describes the behavior of particles at the quantum level (the Hamiltonian dynamics) and the equations

[153] Cf. Bishop, R.C. (2005, 2009).

[154] Cf. Bishop, R.C., Atmanspacher, H. (2006)

[155] Cf. Bishop, R.C., Silberstein, M. (n.d.).

[156] The word "behavior" in the context of physics refers to the changes in the properties and laws that we can use to describe the system under study: "With respect to the behavior of a physical system, we can distinguish the *state* of the system, its *constants*, and the *laws* that pertain to it. Some quantities of a physical system are constant; others vary with time. In the case of a single classical particle, we can distinguish position and momentum as changing quantities, whereas mass remains constant. The values of the varying quantities at a particular time are called the *state* of the physical system at this time. However, the constants and the state of a system do not determine the complete system's behavior. Furthermore, we have laws that describe the connections between the various quantities involved, and in particular, they describe how the state of the system develops in time (the *dynamics* of the system)." [Hütterman, A. (2005), p.115].

used in the field of many-body physics is discontinuous. The behavior before and after that transition is qualitatively different and has to be described by a totally distinct equation. And in order to move from one equation to the other, one has to expand the former about a singular limit, assuming Planck's constant to tend to zero – which is a mathematical trick that departs from reality (in reality, Planck's constant is non-zero). Hence, between the classical and the quantum domains there is a cliff that no bridge laws[157] can cross.

Many other phenomena share this feature of being brought about, in the models, by singular limits. The appearance of molecular shape, the passage from statistical mechanics to thermodynamics and criticality are three such examples that I will now turn to explaining briefly.

The first case has to do with **isomers**, molecules that share identical chemical formulas but have different spatial arrangements which gives them very different properties (one isomer might be lethal for humans while the other is a useful medicine, as in the case of thalidomide). According to Bishop, these are good candidates as examples of contextual emergence since the specific structure into which a certain Hamiltonian will evolve at the chemical level cannot be deduced from quantum mechanical data alone.

> "Even though QM contains necessary conditions in terms of nucleons, electrons and their properties, fundamental force laws and so forth, observables relevant for molecular structure do not exist in the domain of QM. For such observables to obtain, an additional context not given by QM must be specified."[158]

Only with the help of heuristic formal procedures, like assuming the nucleus of the atom to be stationary and infinitely larger than the electron mass[159], can one derive the

[157] Ernest Nagel (see footnote 143) famously introduced the idea that, in cases of heterogeneous reduction (when the terms of the upper-level theory are not a subset of the lower-level theory), bridge laws are required to connect the two levels.

[158] Bishop, R.C. (2009), p.177.

[159] This "clamped-nucleus" assumption is part of the so called Born-Oppenheimer "aproximation". Mathematically, it corresponds to an asymptotic series expansion in which the parameter ε (= electron mass/ nuclear mass) diverges to zero, that is, the nuclear mass is assumed to be infinitely large with respect to the electron mass.

equation encoding molecular shape. This means that the chemical context (the stability conditions of a "clamped nucleus" together with the ratio of the electron mass over the nucleus mass tending to zero) must be fed into the mathematical treatment of the quantum mechanical information. It is thanks to these constraints that come from "outside" the quantum realm that the quantum correlations between nuclei and electrons are broken and classical position and momentum observables, as well as molecular shape, can arise.

The second case has to do with **temperature**. Even though in the philosophical literature temperature is still cited as a good example of reduction (it is taken to be *nothing but* the mean translational kinetic energy of molecules in a system), according to Bishop, it is actually a good example of an emergent feature. Temperature arises out of two mathematical transitions (from particle mechanics to statistical mechanics, and from there to thermodynamics), the calculation of which depends upon mathematical limits (e.g. the thermodynamic limit, which assumes the container of a gas to be infinitely large) as well as on stability conditions which are not available in the underlying domain, such as thermodynamic equilibrium. So again we are before a case in which the appearance of the macro property is not a mere quantitative derivation from a smaller scale to a larger scale, but rather a qualitative transformation which can be explained and predicted (at least on the basis of the models and theories presently available to us) only through the artificial normalization of singular limits. This can be interpreted as an indicator of the inadequacy of our theories and models, or instead as a "source of information"[160]. Robert Batterman has been arguing for the latter attitude for several years now:

> "If it were not for the singularities that appear in our theories and models we would have no understanding of the emergence at different scales of distinct and apparently "protected" states of matter."[161]

[160] Batterman, R.W. (2011), p.1038.

[161] *Idem*, p.1040. The "protected" states of matter that Batterman is referring to are what Laughlin and Pines call "protectorates" [Laughlin, R. B., Pines, D. (2000)], which are stable states of matter that are insensitive to changes at the micro-level, such as in cases of thermodynamic criticality. These protectorates are the units of the phenomenon physicists call "universality", which is what philosophers name "multiple realizability".

According to Batterman, emergence happens precisely there where singular limits cause our lower-level theories to break down. And as a matter of fact, our most important physical theories are asymptotically related in pairs:

$\text{Lim}_{1/c \to 0}$ (special relativity) → Newtonian mechanics

$\text{Lim}_{\lambda \to 0}$ (wave optics) → ray optics

$\text{Lim}_{h \to 0}$ (quantum mechanics) → classical mechanics

One example that Batterman uses in order to show how Nagelian forms of reduction are explanatorily inadequate in cases of singular transitions is thermodynamic **criticality** (the third case listed above). The critical point of a fluid is a state in which liquid and vapor can coexist, and it is determined by a specific temperature and pressure (which is different from fluid to fluid)[162]. Surprisingly, once they reach their specific critical point, all fluids (as well as magnets) behave in an identical manner, even if their properties are radically different in other phases and even if the values of their critical point are as diverse as 1,040.85^{o}C/270 atm for sulfur and -239.95^{o}C/12.8 atm for hydrogen. This macroscopic similarity beyond microscopic differences is what physicists call universality and it has been mathematically accounted for by the renormalization group theory[163]. This mathematical technique (for which Kenneth Wilson won the Nobel Prize) shows how the molecular details that are specific to each fluid are irrelevant for the macroscopic behavior that it shares with all other fluids. The process is based on an iterated transformation of the Hamiltonian of each system, by which as one gradually changes scale, more and more fine-grained information is lost and the resulting function ends up being the same for all the elements of the universality class in question (a value that is called a "fixed point"). Batterman's argument is that "this kind of strategy can provide an explanation for universal/multiply realized behavior without satisfying the criterion of derivability that is essential for Nagelian reduction"[164], and this means that renormalization group theory

[162] As with the previous cases, this phenomenon too is described as the result of assuming a variable to be infinite: viz. the number of particles or the correlation lengths between them (the distance over which one particle can influence another).

[163] Developed by Kadanoff, Fisher, and Wilson [Cf. Batterman, R. (2002)].

[164] Batterman, R.W. (2014), p.15.

challenges the reductionist pretense of providing all the explanatorily relevant information we might need.

Batterman is in good company, as voices have been raising in the attempt to tell philosophers and unexamined reductionists that real-world science does not actually have any models that drill down from many-body physics to some mythic "microphysical state" and that, in fact, productive scientific models largely ignore such thinking altogether. The idea that one might in principle deduce the goings-on in the domains of chemistry, biology and other special sciences from a complete knowledge of particle physics is proven absurd already at the level of condensed-matter physics. As Nobel laureate Robert Laughlin has been defending in the past years, nature is filled with emergent phenomena, regulated by "higher organization principles" and insensitive to microphysics[165]. According to Laughlin and Pines, that the behavior of these phenomena is determined by higher organizing principles is something obvious for solid-state physicists and chemists as well as biologists, even though for other scientists, namely many physicists, this is a dangerous idea that conflicts with reductionism – a belief that is central to much of physical research. However, they say, "the safety that comes from acknowledging only the facts one likes is fundamentally incompatible with science. Sooner or later it must be swept away by the forces of history"[166]. What Batterman adds to this argument is a concrete positive way of explicating these higher principles and how they arise.

3.3. Physics' case against reductionism

What the aforementioned examples show is that the reductionist ideal of macro properties being explainable in terms of microscopic features and laws is not grounded in

[165] Two examples: "The Josephson quantum is exact because of the principle of continuous symmetry breaking. The quantum Hall effect is exact because of localization. Neither of these things can be deduced from microscopics and both are transcendent, in that they would continue to be true and to lead to exact results even if the Theory of Everything were changed" [Laughlin, R.B., Pines, D. (2000), p.261].

[166] Laughlin, R.B., Pines, D. (2000), p.264.

scientific practice. Scale transformations are highly problematic and many aspects of reality seem to simply pop up when a certain threshold of complexity is crossed, which mathematically corresponds to unphysical singular limits. Therefore:

> "Predicting protein functionality or the behavior of the human brain from these equations is patently absurd."[167]

But isn't this a merely epistemic matter? Even if we cannot predict upper-level phenomena on the basis of our lower-level knowledge, this does not imply that we are dealing with ontologically irreducible features.

We face a problem once we start trying to use the epistemological/ontological distinction in physics. Physics does not ever pretend to *know* the reality underlying its models. All physics does is design theories that are quantitative, predictive and falsifiable. Whether those theories correspond to the actual objective truth is something physics cannot tell us. Such an instrumentalist approach can make it hard on the philosopher to extract useful information from physical theory and practice for her metaphysical speculations.

But philosophers do not give up on what might be a fruitful dialogue and try to bridge the two fields in order to inform their theories with scientific information. For example: the philosopher's most typical way of reasoning about the nature of reality in and of itself, independently from our epistemic access to it, is to imagine a universal and omniscient calculator (a laplacean demon), whose complete knowledge of a certain system might be sufficiently explanatory of all its macro properties. Could such a calculator predict the formation of Rayleigh-Bénard convection cells[168] in a fluid with such and such initial and boundary conditions, might the philosopher ask a physicist?

Unfortunately, the physicist will likely find the idea of a laplacean demon to be inapplicable to physical science for several reasons.

---

[167] *Idem*, p.260.

[168] Rayleigh-Bénard convection is a macro phenomenon that is easily observed in liquids that are submitted to a non-uniform temperature distribution (by putting them in a horizontal plane and heating it from bellow), which causes the formation of a regular pattern of geometric cells of moving fluid. Robert Bishop often uses these cells as an example of contextually emergent entities.

First, because if we are dealing with open systems, then the information the demon would have to compute would be the whole universe, and this makes no sense in physics. The models physics works with must be applied to a defined and limited system, which means that this demon would have to work with a different science and different conceptual tools, so to speak; hence, our present physics cannot evaluate its metaphysical possibility any more than can our common logic.

Second, because the *in practice* impossibility of calculating all the information contained in any macro system, not to mention the whole universe, is considered by physicists to be an *in principle* impossibility. It is presently established that no computer can ever accurately solve the quantum Hamiltonian of a system with more than ten particles[169], because the complexity of the equations to be solved grows with the factorial of N (number of particles), which means that the interactions between particles are intractable. So to imagine a universal calculation of the evolution of an ideal "system of the world" just sounds plainly absurd. A philosopher may tend to react to such an argument with dismay and call attention again to the hypothetical nature of the laplacean demon that need not suffer from the physical limitations of actual computers (which store information in space and take time to calculate), but the dialogue with the physicist will likely have come to a halt.

Third, because even if our hypothetical physicist interlocutor decides to take this calculating impossibility as only a practical problem, he will probably present us with one more argument for discharging the laplacean demon from the debate: the theories we have that describe and explain macro properties on the basis of molecular properties are statistical in nature. That is to say they do not express the sort of one-to-one causality relations we would like a laplacean demon to have access to. So a really carefully imagined omniscient being would have to have a theory set that is fully coherent across scales, which is something we are nowhere near to achieving and cannot even know is possible.

All in all, reductionism seems more like a leap of faith than a sound basis for which science has produced any evidence. Every microphysical law, which is an abstract construct

[169] Cf. Laughlin, R.B., Pines, D. (2000), p.160.

formulated by theoreticians in as simple and context-free a way as possible, is tacitly implying that its application depends on the absence of outside influences (influences from upper-levels of organizations). What happens in a laboratory, then, is the testing of such abstract physical laws in equally aseptic environments, carefully designed to exclude any disturbing factor. Like John Dupré says:

> "Very specialized phenomena in extremely carefully controlled conditions do exhibit some impressive regularities (…) produced in extremely elaborate *machines* – machines painstakingly designed for the very purpose of producing these regularities."[170]

However, outside these controlled setups, things get very messy. Even though the results of the experiments often corroborate the laws we wish to test, they cannot confirm their applicability to real-case scenarios where the boundaries between organization levels are loose and causal interactions between them much more likely.

Hence, it would be fallacious to infer from the results of experimental scientific research such strong metaphysical assumptions as reductionism or the causal closure of physics (which I will address in section 3.6), since that would require us to be able to ascertain with profound detail what happens in increasingly complex and ever changing contexts. Evidence for reductionism should consist in the verification that the behavior of complex systems (from chemical, to biological, to neurological, to psychological, to social), in real-case situations, can be fully explained by microphysical laws, which is something that cannot even be done at a molecular level.

In principle reductionism is impossible to prove and so is non-derivability. We can only use as arguments what science is able to verify right now, not what it might be able to reveal in the future. So what both alternatives must do is try to make the case for the higher implausibility of their rival position. In this sense, the epistemological emergence of many-body properties and the radical mathematical discontinuity between theories at different levels does come in handy as evidence in favor of the ontological non-reducibility of the macro to the micro.

[170] Dupré, J. (2001), pp.164-165.

3.4. From anti-reductionism to emergence

But even if phenomena such as molecular shape can reveal the problems facing reductionism, does that amount to what we call ontological emergence? Recall my definition:

> *Ontological emergence is a relation between entities (substances or properties) belonging to different levels of organization of reality (different lengths and time scales), in which the lower-level structure gives rise to the higher-level entity without the novel causal properties manifested at the emergent level being derivable from or causally explainable in terms of the properties of the constituent or subvenient elements that originate them, nor in terms of the laws that govern their behavior and interactions.*

Emergence relies mostly on the irreducibility of the *causal powers* manifested at the higher level. It is not merely a matter of there being novel properties (like chirality) and laws (such as the second law of thermodynamics) that were absent at lower levels of organization and which are not derivable from first principles as Nagel and others supposed they should. It is mainly a matter of the emergent entitiy's new causal power to produce effects in the world not being identical to, nor explainable only on the basis of, the properties of the lower-level entity that brings them about. Upper-level causation (both same-level and downward) is inherent to emergent phenomena, otherwise they are not "emergent" but merely collective phenomena[171].

Liquidity, for example, is a very familiar property (or cluster of properties) which I find it hard to consider emergent, despite its unpredictability and novelty with respect to the properties of the components of the liquid taken in isolation. The macroscopic properties of liquids (like viscosity or surface tension) and their causal powers are entirely explainable in terms of chemical bonds and other microscopic states and events. What we see at the

[171] Even though physicists use "collective" and "emergent" as synonyms, I am taking "collective" to mean systemic (i.e. a property that can exist only at the level of the whole) but explanatorily reducible.

level of the liquid is not something *over and above* the goings-on at the level of the molecules and their interactions. So reducibility seems possible, almost unavoidable. However, condensed matter physicists will tell me, it is not that simple. Note what Victor Weisskopf, one of the founding fathers of quantum mechanics, said about the unpredictability of liquids:

> "Assume that a group of intelligent theoretical physicists had lived in closed buildings from birth such that they never had occasion to see any natural structures. (…) What would they be able to predict from a fundamental knowledge of quantum mechanics? They probably would predict the existence of atoms, of molecules, of solid crystals, both metals and insulators, of gases, but most likely not the existence of liquids."[172]

The stability of liquids depends on temperature, which cannot be derived from the particle's interactions. The calculation of the macro properties of liquids cannot be made without first establishing the stability conditions upon which the liquid depends, that is, one cannot derive their macro properties from their micro state without taking into account the macro conditions that make it so that some laws of interaction rather than others apply. So, why should I consider liquidity, solidity and other ordinary bulk properties any different from the more unusual and paradigmatic properties that Bishop, Batterman and others use as examples of emergence?

Molecular shape, temperature and criticality are considered to be good candidates for emergence because of the irreducibility of their formal description to the underlying Hamiltonian. If this epistemic irreducibility were to express a deeper ontological irreducibility, that would mean that there is a spontaneous and unexplained symmetry breaking at a certain point in the evolution of the system, whereby new properties *with new causal powers* come about. Isomers with different boiling points and densities, temperature with different effects on macroscopic bodies (such as melting), critical points in which new visible phenomena such as opalescence take place (the fluid becomes opaque and colored). If our epistemic limits express true ontological irreducibility, these

[172] Weisskopf, V.F. (1977), p.202.

examples, as well as many others, which might be more or less familiar and more or less complex[173], all seem to be cases of causally new and irreducible, hence emergent, macro features.

However, the new properties we find here are all derivable from the microscopic properties, *given stability or other constraints*. Molecular shape, temperature and criticality are not unexplainable, it we take into account these constraints, and thus are not different from liquidity, which too can be calculated *afterwards* but cannot exist unless there is a certain sort of symmetry breaking induced by temperature. So either *all* of these phenomena can be considered to be emergent, or none of them can.

Let us sum up. These are all systemic properties that are *qualitatively* different from the properties of the parts. They can be calculated once we know the stability conditions that allow them to persist, but the singular limits that separate the theories that describe them render it impossible to explain the whole only on the basis of the parts.

Hence, when authors like John Searle cite properties as ordinary as solidity, liquidity and transparency as cases of emergence, which we would be suspicious of on a first impression, they might actually be right. Not so much because these are "system features [that] cannot be figured out just from the composition of the elements and environmental relations [and rather] have to be explained in terms of the causal interactions among the elements"[174], which is trivial (a clock would fit into this description), but given the fundamental discontinuity between the models that describe the constituents and those that describe the wholes.

Nevertheless, the fact that we cannot know whether this epistemic irreducibility corresponds to ontological irreducibility rather than to mere limitations of our models prevents us from being able to assert whether these are cases of ontological emergence or not. In the end, the move from the epistemic to the ontological level of analysis is a matter of personal preference and intuition. Physics is silent about what is really *there*

[173] Among the more familiar ones we find ferromagnetism; among the less familiar, there is the Bose-Einstein condensate or the quantum Hall effect, which are often used as examples of emergence in the literature.

[174] Cf. Searle, J.R. (1992), p.111.

and so all it can do to help the emergentist's case is tell her that ontological emergence is not an absurd anti-scientific hypothesis. It is actually plausible, if our theories are true, since the way our models relate to each other is exactly what one should expect if ontological emergence were true.

To conclude this reflection on emergent phenomena in physics: Brian MacLauglin, whom I've quoted in the beginning of this chapter, is plainly wrong. Despite the advent of quantum mechanics and other scientific theories, there is much more than a few "scintillas of evidence" that there are emergent causal powers and laws, as well as downward causation, in the world. And they might be much more common than usually supposed, even though reductionism cannot be disproven.

3.5. How is ontological emergence possible?

We cannot be sure on the basis of the examples we collected from condensed matter physics, that there is more than plain epistemological emergence. But there are examples from other fields that might convince us more, especially when we move from the world of material objects located in space-time to the elusive world of conscious mental entities. I will address the possibility of emergence in the transition from the body to the mind in the fourth chapter. Before, however, I need to explore a bit further the conundrums of ontological emergence, should it be true of some of the phenomena we discussed so far, as well as others. The crux of the matter is how the coming about of these irreducible causal powers can possibly take place in the natural world. The appeal to emergent *ad hoc* laws whereby the new feature appears "as a brute fact" makes this all sound too mysterious. Could we give ontological emergence a more detailed explanation?

This is what has led Gil Santos, for example, to argue that for ontological emergence to be conceivable, we have to abandon the atomistic heritage that still makes most of us (classical and contemporary emergentists included) take for granted that the parts are not intrinsically transformed when combined in a whole. In fact, putative emergent properties, in the usual accounts, are properties of the whole, not of the parts; the parts

are considered to undergo only quantitative changes, and to give rise, by mere combinatorial organization and new emergent laws, to different wholes with different properties. That different combinations produce different results is trivial; any reductionist will accept this kind of transformation, which is what allows for complexity and plurality to originate from a few simple principles and a collection of elementary particles. What emergentists add to this upward production of variety is *in principle* unpredictability, that is, the fact that in cases of emergence the structural building up of new systemic features breaks the laws of superposition that otherwise prescribe the outcome of any combinatorial arrangement.

According to Santos, this is something an atomistic framework prevents us from giving a naturalistic explanation to, as Diderot had already noted in 1751, while addressing the quality of living organisms:

> "Life cannot be the result of organization. Take three molecules A, B, C; if they are not alive in the combination A, B, C, then why should they begin to live in the combination B, C, A, or C, A, B? This is inconceivable."[175]

The example of Life can be generalized: mere differences in combination cannot justify the appearance of properties that are simultaneously new (qualitatively novel) and real (not epiphenomenal). That emergence cannot be explained did not seem to disturb its first proponents. However, it should worry those who wish to give this type of account more credibility among naturalistically inclined philosophers and scientists. Santos believes the problem can be solved, for it is grounded in those atomistic assumptions that make the emergence of life become just another word for vitalism: since the components do not change in themselves, the unpredictable novelty we find in the systemic properties of the whole is entirely due to the "new kind of relatedness"[176] that magically transforms the system. This *sui generis* law is not a force (like Bergson's *élan vital*) or a substance (like

[175] "La vie ne peut être le résultat de l'organisation; imaginez les trois molécules, A, B, C; si elles sont sans vie dans la combinaison A, B, C, pourquoi commenceraient-elles à vivre dans la combinaison B, C, A, ou C, A, B? Cela ne se conçoit pas." [Diderot in the *Encyclopédie*, translated by Santos, G.C. (2015b), p.8].

[176] Morgan, C.L. (1923), p.6.

a soul) coming from outside the material realm but it is an *ad hoc* element nonetheless. And this reasoning applies to all sorts of emergent phenomena, not just life, of course.

Instead, Santos suggests that the atomistic model of inalterable fundamental elements arranged in different combinations should be replaced with a relational ontology "according to which all entities' type identities and behaviors are constructed by their intrinsic and extrinsic relations"[177]. For an emergence relation to be possible there must be some qualitative transformation of the intrinsic[178] properties of the constituents of the whole.

Santos' solution is reminiscent of Paul Humphreys' "fusion", by which "when emergence occurs, the lower level property instances go out of existence in producing the higher level emergent instances"[179]. However, while fusion may be an adequate description of what happens in some cases (as in the case of quantum entanglement, which Humphreys uses as an example), it does not apply to many other cases in which the components' parts are still present and identifiable, despite the transformation of their properties. For example, Santos uses the case of symbiogenesis, an evolutionary theory according to which eukaryotes originated from prokaryotes[180] by a process of endosymbiosis[181]. In this process, a host cell (a heterotrophic protist[182]) ingests a cyanobacterium that, through evolution (in which both exchange genes), eventually becomes an organelle, part of a new

[177] Santos, G.C. (2015b), p.18.

[178] By intrinsic properties, Santos means "the properties that an entity has of itself, despite its relations with other entities in its environment – that is, the possession of those properties depends entirely upon what an entity is like in itself – and *relational* properties are properties that an entity has and acquires solely due to its extrinsic relations with other entities" [Santos, G.C. (2015b), p.5, nota 4].

[179] Humphreys, P. (1997), p.8.

[180] Eukaryote: an organism whose cells contain a nucleus and other organelles enclosed within membranes. Procaryote: a single-celled organism that lacks a membrane-bound nucleus and other organelles.

[181] Cf. Santos, G.C, Santos, R. (2012), "Symbiosis as a case of Emergent Evolution". Conference at the seminar "Evolution of Cellular Complexity: Philosophy, Cell Biology and Symbiosis" (November 22nd, 2012. Faculty of Sciences of the University of Lisbon).

[182] Heterotroph: an organism that cannot fix carbon and uses organic carbon for growth. Protist: a unicellular eukaryote organism.

autotrophic organism[183]. The two components are transformed (both the protist and the cyanobacterium have changed properties: the former by becoming autotrophic, the latter by losing its autonomy) and through this transformation a new entity is born: a proto-alga. It does seem too radical to say that there is "fusion" in this case, as the remainder of the two initial components is still there, in the form of the parts of the new whole.

A relational ontology could in fact "solve the mystery" of emergence by revealing that there was no part-whole micro-physicalism[184] to begin with. Things do not exist in isolation and the properties by which we define the objects that exist are always given in a relation. When objects change their relations, their properties change accordingly (I can be a mother only in relation to my children). According to Santos, if that change is qualitative (and not merely quantitative), in that some previously manifested properties disappear and new properties arise, then we are before a case of emergence.

Is this scientifically plausible? It does not seem implausible to me. There are many examples at the level of microphysics that fit a relational model. Elementary particles, for instance, are actually a system of interrelated quarks which can undergo substantial qualitative changes in virtue of the structures they are part of[185]. And quantum entanglement, which I have mentioned before, is usually taken as the quintessential example of a system in which the parts' properties are correlated and dependent on the compound system as a whole.

However, if one takes emergence to be a relation by which an entity acquires independent causal powers that are not derivable from its emergence base (and what relevance could this concept have if we endowed it with any less autonomy?), I cannot see that Santos' solution is helpful. The causal properties of the emergent whole (e.g. the eukaryote) are directly derived from the properties of its *transformed* parts (e.g. the protist and the cyanobacterium). And that transformation seems quite linear to me: the gene exchange

[183] Autotroph: an organism capable of producing complex organic compounds from simple substances using energy from light (photosynthesis) or inorganic chemical reactions (chemosynthesis).

[184] Cf. the definition of micro-physicalism by Andreas Hütterman (2005).

[185] Cf. Cordovil, J.L. (2015).

and other mechanistic processes perfectly explain how the protist became an organelle and how the original heterotroph became an autotroph. There is no more novelty here than in cases where classical reductionism applies, i.e. when the components, their properties and the laws that describe them, plus the laws of composition that govern their addictive relations, are sufficient to determine the outcome. In contrast with those cases, the concept of emergence implies the possession of *basic* causal powers by the upper-level entity, despite the apparent contradiction of an emergent feature being basic. So the case of the new autotroph organism is not a case of ontological emergence at all, according to this definition.

Santos might counter that I am begging the question. I am adopting a much too strong concept of emergence, which will naturally exclude the cases he is concerned with. He could suggest that I might instead accept that there are weaker and stronger versions of ontological emergence[186]: the former would regard cases like endosymbiosis in which the properties of the whole are brought about by a process of qualitative transformation of the parts, without that implying that the causal powers those systemic properties carry with them are unexplainable in terms of the new transformed properties of the parts; the latter would concern cases where the emergent features would have new basic causal powers. Even if the former cases would fail to meet the criteria for strong ontological emergence, they would be emergent nonetheless, if we define emergence against the background of part-whole micro-physicalism.

True, one could define emergence in such terms, but that would be a much less familiar definition and one that renders secondary the feature that in my opinion (and in that of many authors) is the crucial element for drawing the line between emergent and non-emergent entities: novel causal powers.

Let us consult the Stanford Encyclopedia of Philosophy, for instance, to see how emergent properties are most commonly defined:

[186] Note that the distinction between epistemological and ontological emergence is very often made in terms of "weak" and "strong" emergence. What I am supposing might be objected to me in terms of a useful distinction between strong and weak versions of emergence would instead regard only cases of ontological emergence.

> "It is a novel, fundamental type of property altogether. We might say that it is 'nonstructural,' in that the occurrence of the property is not in any sense constituted by the occurrence of more fundamental properties and relations of the object's parts. Further, newness of property, in this sense, entails new primitive causal powers, reflected in laws which connect complex physical structures to the emergent features."[187]

And what are new causal powers usually taken to be in the literature? Let us see how Jaegwon Kim defines them:

> "[E]mergent properties (…) are supposed to represent novel additions to the ontology of the world, and this could be so only if they bring with them genuinely new causal powers; that is, they must be capable of making novel causal contributions that go beyond the causal powers of the lower-level basal conditions from which they emerge."[188]

What Santos classifies as emergent are systemic properties which arise through a process of transformation of the properties of that system's parts. This makes it so that between the macro level of the system and the micro level of the parts as taken in isolation, a new intermediate *meso* level of the transformative relations between the parts must be taken into consideration. I find this type of process trivial. The occurrence of properties which arise out of these transformative relations *is* "constituted by the occurrence of more fundamental properties and relations of the object's parts" and it is *not* "capable of making novel causal contributions" that go beyond them.

In the case of endosymbiosis, used as an example above, despite the natural evolution whereby these two entities lost some of their properties and became a new type of entity with novel causal powers (such as the ability to produce its own food through photosynthesis), one can describe what happens whenever this entity exercises its new powers in micro chemical terms, with no loss of information. So even if diachronically, in

[187] O'Connor, Timothy and Wong, Hong Yu, "Emergent Properties", *The Stanford Encyclopedia of Philosophy* (Spring 2012 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2012/entries/properties-emergent/>.

[188] Kim, J. (1999), p.25.

the long evolutionary time, there was a qualitative macro transformation, synchronically, reduction is still the best explanation of the fine-grained mechanisms of a proto-alga's photosynthesis.

This is not to say that diachronic emergence is not important. The point is just that one cannot claim that the original coming to be of a new entity at a certain point of evolution is a case of ontological emergence if, synchronically, the subvenient parts of the composite system, together with the laws that govern their relations, are sufficient to derive the upper-level properties in question.

Let me be clearer: I take diachronic emergence to be a causal relation between entities, which involves a time lapse between an instant t when the emergent entity was still absent from reality and the moment t_1 when it first appears. It may concern entities belonging to different domains (e.g. physical vs. chemical) or same-level entities. In turn, synchronic emergence is a realization relation[189] between simultaneous entities at different domains. Both these versions of emergence can be epistemological or ontological and both involve the crucial discontinuity that founds the irreducibility of the emergent to its basis.

On my view, there can be no diachronic emergence without synchronic emergence, nor vice-versa[190]. If the putative diachronically emergent entity is constituted by its parts or realization base in such a way that one can identify all the steps that lead from the lower-levels to the upper-level phenomenon, then the diachronic process that causally produces this entity over time is similarly explainable. How can we say there was the intervention of a *sui generis* law by which new causal powers (irreducible to the causal powers of its diachronic causes) appeared, if the lower-level elements are sufficiently explanatory of upper-level phenomena? If they are sufficient now, they would have been sufficient in the past as well.

[189] I take realization to be an umbrella term for inter-level compositional relations (between parts and wholes) as well as relations between entities of different nature belonging to the same level of organization (such as neurobiological and psychological entities).

[190] Olivier Sartanaer argues for this same thesis in his (2015).

To sum up, I think the symbiogenesis example is not a case of emergence. Also, it seems to me that to assume a relational rather than an atomistic ontology is not what makes the greater difference to the question whether ontological emergence exists and can fit the natural world. Even if one adopts a relational ontology according to which the components of a system can undergo qualitative changes in their intrinsic properties, the new features that appear as a consequence of such transformations should be considered emergent only if their causal properties are basic, which entails that they are underivable from the initial properties of the components (both intrinsic and relational) and the laws that govern their interactions. There is no reason why a relational ontology should be more favorable to this type of irreducibility than the traditional atomistic ontology.

In my view, what does make a difference to the plausibility of emergence is whether our world view allows for the influence non-fundamental entities have on fundamental events to be irreducible to the influence bottom-level events have on each other. This presupposes that the entities we find at the most fundamental level of reality are not the only type of entities that have effects in the world, which is what I will discuss in the next section.

3.6. Is the physical causally closed?

The main reason why the emergentist hypothesis is met with suspicion in scientifically inclined circles is the idea that the physical world is closed to causal influences from outside its frontiers. The principle of the Causal Closure of the Physical (CoP), that many mistake for an unquestionable axiom of physical theory[191], asserts roughly that *for every physical event (or its chances*[192]*) there is a sufficient physical cause, insofar as there is any*

[191] Cf.: "Physics does not admit that physical effects have non-physical causes" [Sturgeon, S. (1998), "Physicalism and Overdetermination", *Mind* 107: p. 413; cit. in Lowe, E.J. (2008)].

[192] It is worth noting that contemporary authors tend to substitute the nondeterministic version of CoP in terms of every physical event having *its chances* fixed by sufficient physical causes, for the simpler to state but questionably deterministic formulation in terms of the *physical events* themselves having sufficient physical causes [Cf. Lowe, E.J. (2008), p.44].

cause at all. At face value, this seems to entail one of two consequences: either non-physical properties and objects are actually physical after all, or they are epiphenomenal. In other words, either the chemical, the biological, the psychological and the social domains are only different levels of description of one same reality that can in principle be understood in terms of the laws of a complete physical science, or the putative emergent levels are distinct from the physical reality but there is nothing left for them to cause, since all effective causes are physical causes.

CoP can be read in two different ways: as stating the completeness of the microphysical (as opposed to the chemical, biological, etc.) or affirming the completeness of the physical, intended as the material reality that exists in space-time and can be described or explained in quantitative terms (as opposed to the subjective nature of conscious thoughts, say). The former reading is with no doubt the most common in the philosophy of science literature. The latter is mostly discussed in philosophy of mind, where CoP appears as particularly threatening, since people cherish the causal autonomy of mental properties more than that of any other property. If the solidity, mass and speed of a cannonball are nothing but the global effect of the properties of its tiny parts, that does not disturb us much. But if our conscious states (intentions, desires, decisions) are epiphenomenal, then our vision of who we are and how we act in this world is in need of a major revision.

3.6.1. The Causal Closure of the Microphysical as a typicality condition

Robert Bishop has devoted much of his work to assessing the Causal Closure (or Causal Completeness) principle in the first sense (the “physical” intended as the “microphysical”) He notes that CoP is presented usually as a premise in arguments in favor of physicalism – the thesis that everything is physical or supervenes on the physical. In such arguments,

as for example Papineau's Causal Argument for Physicalism[193], Causal Closure is formulated as the claim that "all physical effects are fully determined by fundamental laws and prior physical events"[194] and it is illegitimately assumed as a well-founded scientific decree of some sort. That, however, is false:

> "Physics itself does not imply its own causal closure nor is there any proof within physics of its own completeness."[195]

Indeed, Bishop argues that evidence from physics supports only a qualified reading of CoP as a typicality condition (stating what happens in scientific labs, under controlled circumstances that prevent non-physical interferences). According to this reading, what CoP tells us is that *in the absence of non-physical influences*, physical causes (events and laws) will produce physical effects. In order for CoP to entail the ineffectiveness of non-physical causes in the etiology of physical effects, one would have to endorse the extra clause "that the only efficacious states and causes are physical ones"[196] – which is clearly a question-begging assumption[197].

Hence, the reductionist physicalist is left with two unattractive alternatives: on the one hand, she can endorse a strong but unjustified interpretation of CoP, which would prevent any emergent entities from having autonomous downward causal powers (preventing

[193] Papineau's argument is the following: 1) All physical effects are fully determined by fundamental laws and prior physical events (CoP); 2) Some mental events are causes of physical events; 3) Physical effects of mental causes are not, in general, causally overdetermined (Causal Exclusion); 4) Mental causes are identical to physical causes (Physicalism). [Cf. Papineau, D. (2002)].

[194] Bishop, R.C. (2006), p.45.

[195] *Ibidem*.

[196] *Idem*, p.47.

[197] E.J.Lowe, who discusses CoP mostly in the second sense (of physical vs. mental reality) presents a similar thesis in (2008, part I) and he too argues that the stronger version of physicalism is actually an "unwarranted dogma" based on the faith in the empirically unfounded claim that "no physical event has a non-physical cause" (p.40). He also says that the weaker versions of CoP are much more plausible than the stronger ones (which either render the physicalist argument question-begging or lack empirical support), and they allow for non-physical causes to enter the causal chain, either by directly causing a physical event (along with its physical causes, which are physically sufficient but would have been coincidental, were it not for the interference of the mental element) or by causing it to be the case that certain physical events have a certain physical effect.

also any intentional action from taking place, according to the argument developed in the previous chapter), but which cannot be confirmed nor disproven by empirical means. On the other, she can adopt the weaker typicality version, supported by physical science, but which is insufficient to ensure that only physical causes are effective. Either way seems to get reductionism into more trouble than anticipated.

In order to defend the plausibility of his weak interpretation of CoP, Bishop uses several examples that show how scientific laws and forces in physics are *ceteris paribus* clauses, by which he means that they are never context-free: for example, the gravitational law can predict the movement of a certain body only if no other forces (e.g. electromagnetic) are affecting its behavior; the half-life of neutrons is dependent on whether they are bound in a nucleus rather that isolated; and the expression of genes depends on the presence of other genes and environmental influences.

> "The upshot of these examples is that all of the forces and laws we take to be important in our sciences always carry tacit clauses of the form 'If nothing outside affects the object, then ...' where 'outside' can be understood as outside the relevant body of theory (other senses of 'outside' more relevant to our concerns here would turn on the construals of 'physical' and '*non*-physical'). In other words, context matters at least as much as laws."[198]

Bishop then concludes that "the question of overdetermination is a contextual affair"[199], because the structure of reality is such that entities are simultaneously influenced by many forces and other bodies, and causation is mainly a cooperative process. Whether a certain cause is sufficient or not for its effect, then, is a matter of context. Even if a force *would have been* sufficient to produce a certain effect in an isolated context, when other entities enter the arena, the game changes and the *ceteris paribus* character of the physical laws describing that force and its correlations is revealed.

The unjustified tendency to accept Causal Closure as a proven natural truth is what mainly feeds the accusation of scientific implausibility that faces emergentism and agent-

[198] Bishop, R.C. (2006), pp.49-50.
[199] *Idem*, p.48.

causalism. However, once we realize that the dispute is clearly a philosophical one, since we have no grounds for taking CoP as anything more than a typicality condition, emergence becomes a plausible explanation for many natural phenomena in which control seems clearly to be exercised in a top-down manner.

3.6.2. Causal Closure of the "Material" world and two types of supervenience

It is important to note that one can challenge the causal argument for physicalism by questioning only the second interpretation of the Causal Closure Principle (which intends the physical in a broader sense – let us call it "material"), and leaving the first untouched. By this I mean that one can consider that the whole material world can be understood in reductionistic terms, i.e., that all the hierarchical levels, from the physical to the biological, are connected by some sort of "building relation"[200] and that any material event has its chances determined by a sufficient material cause. Still the mental world can be seen as something new, distinct and genuinely emergent, something that is ontologically irreducible to the material levels and that has new non-derivative causal powers.

John Dupré is one of the authors that questions CoP in the material-to-mental sense[201], based on the argument that accepting the postulate of causal closure would render mental causation impossible. He uses the example of a person wanting to drink a glass of water (mental domain) and moving her arm accordingly (physical domain): if one assumes CoP to be true, the extraordinary coincidence of the co-occurrence of appropriate events at the two levels (the level of the mental intention to drink and the level of the physical causes than produce the physical movement) remains a mystery:

> "All the physical movements of the agent would have happened even if the mental occurrences had not, if, that is to say, there had been no principles or laws requiring mental processes or events to come along for the ride with the physical ones. (…) Causal

[200] I am borrowing this phrase from Bennett, K. (2011).

[201] Even though he is an anti-reductionist is the first sense as well.

> completeness at the microlevel must entail reductionism, at the very least in the sense of the supervenience of everything else on the microphysical. And even supervenience, I claim, is sufficient to deny any real causal autonomy to higher structural levels."[202]

At face value, in fact, supervenience seems not to leave room for downward causation, as this relation entails that mental properties depend on neurobiological properties, and not the opposite. When a certain brain state at t_1 causes another brain state at t_2, there seems not to be anything left for the related mental state to cause, since all the changes taking place at that mental level (thus, the changes from the mental state at t_1 to the mental state at t_2) are taken to depend on the underlying physical changes[203].

Dupré is not alone in considering that supervenience is an unfounded assumption that conflicts with the causal efficacy of entities at upper levels of organization, most notably at the mental level. Paul Humphreys[204], for example, opposes emergence and supervenience as two incompatible forms of inter-level relation: you cannot have one without giving up the other.

I agree with Dupré that Causal Closure entails the epiphenomenality of any state or event which is distinct from the physical states and events on which it supervenes, and that this is extremely implausible, as will become clear in the next chapter. But unlike him, Humphreys and many others, I believe supervenience is not the problem. At least not natural supervenience.

In his famous 1996 book *The Conscious Mind*, David Chalmers presents a very useful distinction between logical and natural supervenience. *Logical* supervenience is a relation that holds between properties A and B such that there cannot be a world in which A (the subvenient property) is present without B (the supervenient property) being present as well. This is the type of relation that holds between the aesthetic properties of a painting

[202] Dupré, J. (2001), pp.161, 162.

[203] The alternative, of course, would be to assume an identity theory according to which mental events are actually physical events, but that is something Dupré is not willing to do. I will address the reasons why I believe that conscious mental states and events cannot be identical to physical states in the next chapter. For now, I wish only to analyze Dupré's argument as to why Causal Closure entails the causal redundancy of the non-physical.

[204] Humphreys, P. (1997).

and the physical substrate that realizes it. If one wanted to reproduce a certain painting, one would only have to duplicate every single physical particle that constitutes it, in its proper place and entertaining the same relations to all the others as in the original painting, and the aesthetic properties of the painting would come along as a bonus. To use Saul Kripke's[205] famous image, if everything there is logically supervenes on the physical, then when God created the world, he needed only one day to do it. After having created the physical, God could rest, because every single aspect of the world as we know it (all the esthetic properties of every painting, the biological properties of living systems, every thought or phenomenal feel people experience) was in place.

This is the type of supervenience that Chalmers' zombie hypothesis stands against and I will address it in the next chapter. What interests me now is *natural* supervenience, the type of covariance relation that science assumes to hold between physical and non-physical properties (i.e. chemical, biological, psychological properties) in *this* world. This is the type of supervenience that I contend does not have to be called into question even if emergence is true of some phenomena or levels of organization.

Assuming that ontological emergence is possible and that at a certain point in the scale of complexity some emergent entities appear, one does not have to imagine that those entities' autonomy is such that some of the changes they undergo at the upper level might happen regardless of corresponding changes at the bottom level. Even if supervenience holds between the emergent level and its emergence base, such that there is always covariance with a bottom-up entailment, the upper-level entities which are realized at t_1 by the subvenient entities can affect the world at t_2 in ways that are not merely the consequence of how some subvenient material elements affect each other. There is nothing *a priori* incoherent about this, and in the next two sections I will show why.

The crucial step in the argument is that the break of Causal Completeness is not sufficient for emergence to be possible: there is another condition that must be in place, namely fundamental indeterminism.

[205] Kripke, S. (1972), pp.153-154.

3.7. Indeterminism at the bottom

According to Kim's famous Supervenience Argument[206] (often called the Causal Exclusion argument), if mental properties are supervenient on physical properties and yet irreducible to them (in the sense of not being identical to any physical item), then they are bound to be causally impotent[207]. Kim's argument for the inconsistency of the three theses of non-reductive physicalism – supervenience, irreducibility and causal efficacy – is based on what he call's Edwards' dictum[208], according to which vertical determination excludes horizontal causation. He cites the following passage by Jonathan Edwards:

> "The *images* of things in a glass, as we keep our eyes upon them, seem to remain precisely the same, with a continuing, perfect identity. But it is known to be otherwise. Philosophers well know that these images are constantly renewed, by the impression and reflection of *new* rays of light; so that the image impressed by the former rays is constantly vanishing, and a *new* image is impressed by *new* rays every moment, both on the glass and on the eye… And the new images put on *immediately* or *instantly* do not make them the same, any more than if it were done with the intermission of an *hour* or a *day*. The image that exists at this moment is not at all derived from the image which existed at the last predicting moment."[209]

It is easy to see the analogy between this case and the mind-body relationship. If mental occurrences are synchronically dependent on physical occurrences (as mirror images on

[206] A request to my readers is in order: the fact that I am citing Kim's Supervenience Argument should not be misinterpreted. I am concerned with the possibility of downward causation, independently from where in nature it might take place (between the chemical and the physical, the biological and the chemical, the psychological and the biological), so my aim is not to address the mind-brain relation nor any of its conundrums at this point. However, I will use one the most famous arguments in Philosophy of Mind against non-reductive physicalism in order to prove my point. Therefore, I ask my readers to keep in mind that the context in which that argument was first presented by Kim is very different from the context in which I am using it now. For instance, I am assuming by hypothesis that there is no type nor token identity between the emergent entities and their substrate, because I am addressing the question of the conditions of possibility for any emergence relation which, by definition, can hold only between distinct entities.

[207] Cf. Kim, J. (1993).

[208] Named after the 18th century philosopher Jonathan Edwards.

[209] Edwards, J. (1758) cit. In Kim, J. (2003), p.154.

rays of light), their diachronic dependence on putative causal antecedents breaks down. Counterfactually, what happened at t_1 is irrelevant for the persistence of an image (or a mental state) at t_2, in case the supervenience base, which is indispensable, fails.

Kim's way of developing Edward's insight comes in two stages. First, he shows how, given supervenience, mental-to-mental causation entails mental-to-physical causation: the mental event M_1 at t_1 can cause another mental event M_2 at t_2 only by causing the subvenient physical event P_2 on which M_2 depends. And this can be generalized to any causal relation between upper-level properties bound by supervenience to a more fundamental level. Second, he argues that, upon analysis, mental-to-mental and mental-to-physical causation (or any other kind of causal relation involving an upper-level cause) are explained away by the presence of two more theses that physicalists will mostly likely be willing to embrace: causal closure, which I addressed in the previous section, and causal exclusion, according to which "no single event can have more than one sufficient cause occurring at any given time – unless it is a genuine case of causal overdetermination"[210].

If M_2 has to have a physical cause P (causal closure) and it is usually not overdetermined (causal exclusion), and if M_1 is not identical to any physical event at t_1 (irreducibility), then M_1 cannot be the cause of P_2; P_1 most probably is. This entails the epiphenomenality of the mental:

> "In this picture, there is but one causal relation, from physical property P to another physical property [P_2], and the initially posited causal relation from [M_1] to [M_2] has vanished altogether. An apparent causal relation between the two mental properties is explained away by their respective supervenience on two physical properties that are connected by a genuine causal process."[211]

So "either reduction or causal impotence", Kim says. And I would follow him on this, if I accepted all of his premises. But I do not. I question causal closure (which Kim, instead,

[210] Kim, J. (2003), p.157. This is basically thesis 3 of the Papineau's Causal Argument for Physicalism, which I addressed in the previous section.
[211] *Idem*, p.159.

considers to be "virtually an analytic truth"[212]), for the reasons stated above, and I question also one hidden premise that I think is crucial for his argument: the assumption the world works deterministically.

"No single event can have more than one sufficient cause occurring at any given time", Kim says. So bottom-level causes render upper-level causes redundant because they are *sufficient* to ensure the effect. In fact, it seems logical to me too that, if every detail of the lower-level is already defined, there is nothing left for the upper level to cause, neither in a top-down way, nor as a same-level causal event. However, if one questions Causal Closure and assumes there is quantum indeterminacy at the bottom-most level (as mainstream interpretations of quantum formalism do), then there is room for causal collaboration between different-level causes. I will now argue for this thesis, and it might be useful to start with the help of the following diagrams:

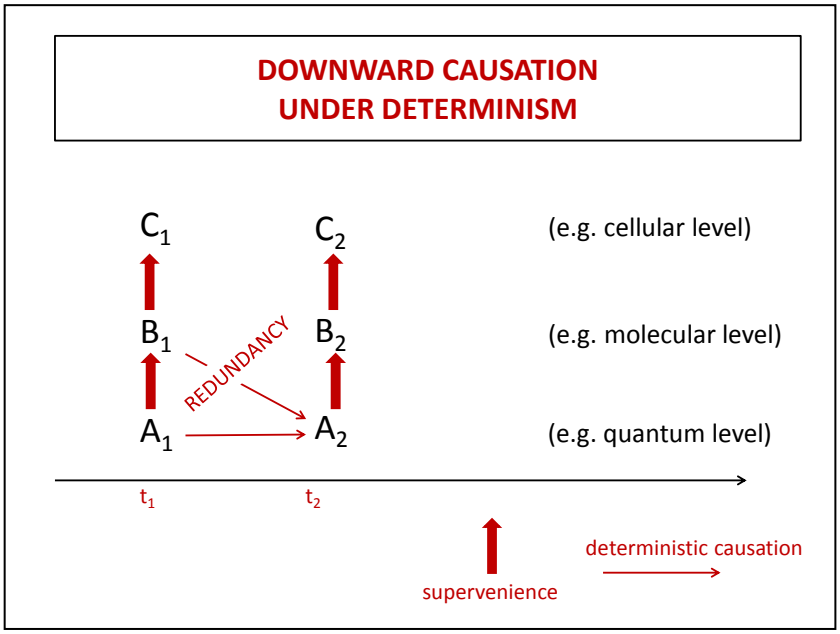


Fig.1: In a world in which the upper levels of complexity supervene on the lower ones, if determinism connects the most fundamental state of a system at t_1 with its following state at t_2, then there is nothing left for any other level to cause.

[212] *Idem*, p.162.

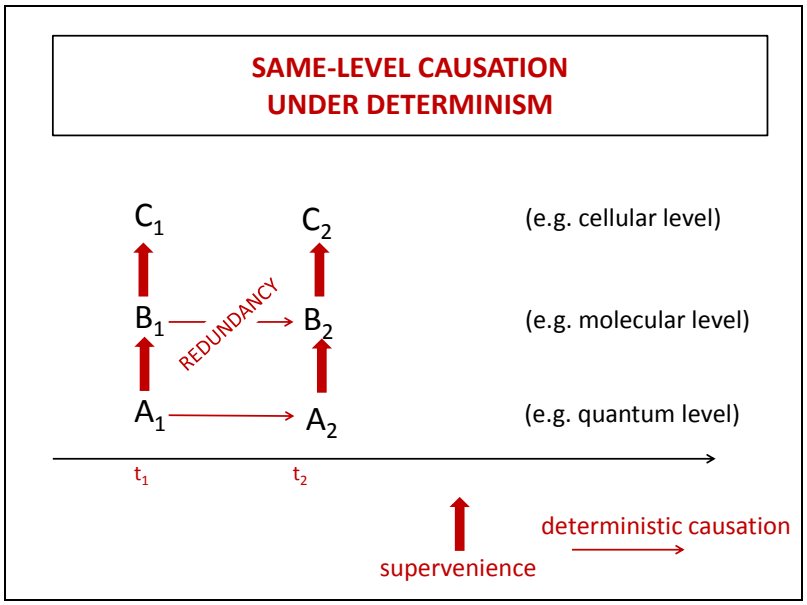


Fig.2: As in Fig.1, there is no room for causal difference-making at t_2 if A_2 is the only possible state of the system at the most fundamental level, and B_2 supervenes on A_2.

Now imagine a dust particle suspended in a fluid, undergoing Brownian motion. Its space of physical possibilities would obviously be much broader if it were outside this context. However, given its location and its relation with all the other particles around it, that possibility space is limited to what the laws allow in the concrete situation it is in. Do Newtonian laws allow for diverse alternatives in the particle's next move given this situation? Of course not, since they are deterministic! Now picture an elementary particle inside your brain, instead. How can neurons and molecules that are made up of billions of particles like this have a downward influence over the particles' movement? How can mental states exercise their causal power over it? Only by means of the interactions each particle has with its neighbors, which constrain its motion just like a dust particle in a fluid, but that happens at the physical (not the chemical) level. The downward constraint exerted by the molecules and neurons that particle is embedded in is only apparent and derivative, for all the determination happens at the level of the parts.

This determination goes back all the way to the Big Bang, as strange as that may seem. If fundamental laws, which describe the rules of interaction among fundamental particles and fields, were universally deterministic (in the sense of allowing for a "unique evolution"[213] of each system and of the world), the initial conditions of the universe as a

[213] Bishop, R.C. (2011), p. 86. The also note 108 in the presente dissertation.

whole (which would include the details of the environment of every given system) would be sufficient to determine its multilevel evolution in time. The causal powers of all the complex macro-structures we find in our 21st century world and their epistemically emergent features and criteria for downward constraint – such as the evolutionary fitness of a cell or the effectiveness of desires in cases of human agency – would not only be derivative and drain down to the elementary basis of the world here and now, but they would have been determined to be what they are and to cause what they cause from the very beginning of the universe.

Let us use John Conway's "Life" game[214] as a model of a deterministic universe. In this model, from very simple two-type individuals and three rules of necessitation, upper level configurations and regularities arise in such a way that we consider them to be epistemically emergent. By this, I mean that the behavior of the macro aggregates and patterns that appear at a certain point in evolution seems unexpected given only the three micro-level rules with which the game begins and, after a certain degree of complexity, it is better described and understood by coarse-grained laws. However, this does not mean that the three "fundamental laws" are insufficient for the existence of those higher level entities. On the contrary, what surprises us in Conway's "Life" is precisely that nothing further had to be added for this new epistemic level to arise, that the necessary and sufficient conditions for such a complex evolution are so simple and were given from the beginning.

Authors who consider that the break of Causal Closure is all that is needed for emergence to be possible, would certainly object that the game of "Life" is not a good example, for it describes a causally closed world, a world in which the fundamental level is complete,

[214] Cf. Gardner, M. (1970): "The basic idea is to start with a simple configuration of counters (organisms), one to a cell, then observe how it changes as you apply Conway's "genetic laws" for births, deaths, and survivals. (…) First note that each cell of the checkerboard (assumed to be an infinite plane) has eight neighboring cells, four adjacent orthogonally, four adjacent diagonally. The rules are: 1. Survivals. Every counter with two or three neighboring counters survives for the next generation. 2. Deaths. Each counter with four or more neighbors dies (is removed) from overpopulation. Every counter with one neighbor or none dies from isolation. 3. Births. Each empty cell adjacent to exactly three neighbors - no more, no fewer - is a birth cell. A counter is placed on it at the next move."

which begs the question. On the contrary, our world is such that the physical domain provides only necessary but insufficient conditions for the other levels to unfold. This is why Robert Bishop, for example, believes emergent phenomena can happen even in deterministic contexts. If the principle of the Causal Closure of the Physical (intended as the Microphysical) is qualifies as a mere typicality condition, determinism becomes irrelevant[215], he says. Once we understand that there is no reason to assume that only microphysical causes can be effective, we can perfectly accept that chemical or biological constraints can influence the unfolding of the causal history of reality.

In contrast, and to use Bishop's own terms, I believe that the break of Causal Closure is a necessary but insufficient condition for there to be non-derivative upper-level causation in the natural world, and thus for ontological emergence to be effective. Even if we accept the interference of non-fundamental (or non-material) entities in the lower-level sequences of events as *nomologically* possible, there will not be any *logical* space for them to be difference-makers if causal interactions at the lower-level are deterministic.

George Ellis, a leading cosmologist who has recently been working on the philosophical problem of emergence, talks about top-down causation in the human brain as a consequence of the break of Causal Closure:

> "Physics provides necessary conditions (but not the sufficient conditions) for what happens; it provides the possibility space for what happens, but does not determine the outcome. Top-down causation allows higher-level causes to be what they appear to be: real effective causes. Context is the key to physical outcomes: multiple causation is always at work."[216]

In other words: when the biological system which we call an agent decides to make a bodily movement such as waving at a friend, she is influenced by reasons, motives and traits of character, just as she is causally conditioned by deterministic and probabilistic laws, as well as physical impossibilities. The complex interaction among all these elements

[215] Cf. Bishop, R.C. (2010).

[216] Ellis, G.F.R. (2009), p.78. It is interesting that Ellis cites Bishop and Atmanspacher's 2006 article explicitly, when referring to the importance of contexts.

is what ultimately structures her possibility space. Also, since Ellis assumes a nonreductive physicalist stance, according to which the upper-level influences are real and distinct from lower-level ones – in short, that those reasons, motives and traits of character are neither identical to their neural correlates nor epiphenomenal.

However, if all the abstract degrees of freedom of each particle in the agent's waving hand are limited by the specific circumstance in which it is embedded (the quantitative values of its mass, charge, location and momentum as well as the values of all the other elements in its physical context), then whatever may be the psychophysical interactions that might take place, there are no more open alternatives than the ones already given, like in the case of the particle in a fluid that we mentioned before. If fundamental physics is deterministic, the trajectory of the particle cannot change, for example, neither can the timing of a neural spike.

If the waving example were a case of emergence, the agent's conscious states would have to be taken as distinct from her neural states as well as endowed with some downward causal power over them. This entails that, when the agent purposefully waves at a friend, for her intention to effectively cause her behavior, it must be the case that *some* particles in her motor cortex have different possibilities of movement at the instant immediately following her decision to move, even given all the specifications of their circumstance at the instant of her decision (the precise values of their mass, charge, location, etc. as well as the complete state of each of their neighbor particles); only if this is so, can they initiate a causal chain that will lead all the other particles correlated with them, namely the ones which constitute the agent's hand, to move according to her will. There must be diverse *possibilia* that may or may not become *actualia* by downward constraint.

Ellis is one of the very few authors who explicitly affirm that indeterminism is a necessary condition for this interaction between levels to take place – even though he does not develop this issue much. Right after the above citation, he adds:

> "Random fluctuations along with quantum uncertainty provide the freedom at the bottom needed to allow this to happen. It enables the causal power of abstract entities – mathematics, theories, ethics, social constructs (…)."[217]

According to Ellis, then, the very possibility of downward causation depends on the existence of genuinely probabilistic laws at the most fundamental microphysical domain - what he calls a "causal slack"[218]. But surprisingly, this demand of indeterminism at the bottom-most level is something that authors seldom mention as a precondition for emergence[219].

Arguing for the effectiveness of mental causation, John Sperry[220] has famously coined the example of a wheel rolling downhill in which each molecule's movement is determined in space and time by the overall properties and dynamics of the wheel as a whole. Even though Sperry's interpretation of such a case as an example of emergence and its use for an analogy with the mutual interdependence between consciousness (the rolling wheel) and the brain (whose individual neurons are "carried along" just like the molecules in the wheel) have both been generally criticized in the literature, I believe this example is useful for the point I wish to make.

Sperry was one of the first authors in contemporary philosophy of mind and, especially, in the world of neuroscience, to defend the idea that conscious mental states were not

---

[217] *Ibidem*.

[218] *Idem*, p.74. One must be aware, though, that indeterminism does not imply the lack of causation. Ellis' phrase is quite imprecise.

[219] Helen Steward is one exception: according to her, if determinism were true, there would be "no gap into which a phenomenon like top–down causation might be fitted" (2014, p. 240). However, my view departs from hers in that she considers universal probabilistic laws also to be a problem for downward causation.

Another exception is Peter Ulrich Tse, a neuroscientist who was part of the four million dollar project "Big Questions in Free Will", sponsored by the Templeton Foundation and directed by Alfred Mele. This project aimed at putting philosophers, scientists and theologians working together on the problem of free will. Tse wrote a book called *The Neural Basis of Free Will* (MIT Press, 2003), where he put forward an empirical model of how downward causation might happen in the brain, allowing for free will. One of the conceptual steps in his argument is that Kim's Exclusion Argument can be overcome if ontological indeterminism is true of neural processes (pp.123-127).

[220] Sperry, R.W. (1969).

only causally efficacious, but were actually so in their own right, and not merely insofar as they were identical to their neural correlates.

> "Conscious phenomena as emergent functional properties of brain processing exert an active control role as causal determinants in shaping the flow patterns of cerebral excitation. Once generated from neural events, the higher order mental patterns and programs have their own subjective qualities and progress, operate and interact by their own causal laws and principles which are different from and cannot be reduced to those of neurophysiology."[221]

However, Sperry defended micro and macro determinism in the neural substrate and functioning of the brain, and also argued that, in exerting their "supervenient downward control", the emergent mental properties could not intervene nor disrupt the causal activity at the lower-level[222]. Consequently, the analogy he used for the reciprocal interaction and determination between mental and neural levels had to be that of the trivial but only apparent mutual influences between the micromolecular level of the atoms in the rolling wheel (which are obviously causally efficacious at their own level), and the determination of their space-time trajectories by the entity as a whole (which is actually just a macro level of description that does not correspond to any real superior and novel causal influence over the wheel's components).

I believe Sperry's use of the example of the wheel is particularly significant in the context of his explicit defense of the compatibility of emergence and universal determinism, insofar as it shows precisely how the Newtonian laws that govern or describe the behavior of physical entities do not leave any room for downward causation. What does it mean for atoms and molecules to be "carried along" as the wheel rolls downhill? The rolling of the wheel itself is nothing but the sum of the movements of its atoms and molecules, which lower level laws manage perfectly to describe within the framework of a reductionist physics.

[221] Sperry, R.W. (1980), p.201.

[222] These two theses, so I argue, are interdependent, thus I believe Sperry's contention of the latter (that emergent laws do not contradict the lower-level ones) was actually a coherent consequence of his adopting the former (determinism).

It is interesting to see how Sperry felt the need to explicitly state the differences between his thesis and Popper and Eccles' dualistic theory of the mind-brain relation in his 1980 commentary "Mind-brain interaction: mentalism, yes; dualism, no"[223], of which a specific section is dedicated to the debate of "determinism versus indeterminism". In it, Sperry defends the deterministic functioning of the brain throughout all layers, fundamental and emergent, as well as the causal determination of the synchronic emergence process itself[224], while Popper and Eccles, in their influential book, had defended precisely what I believe to be true:

> "The emergence of hierarchical levels or layers, and of an interaction between them, depends upon a fundamental indeterminism of the physical universe."[225]

But maybe the example of the wheel was just unfortunate. Could a better instance of deterministic emergence help? Bishop often uses Rayleigh-Bénard convection cells as a paradigmatic example of complex systems in which sensitive dependence allows for the emergence of higher level structures with downward causal power - in other words, a "control hierarchy". However, even though an analogy between this type of system and the human brain would be more accurate than the preceding example because of its increased complexity, still the movement of each molecule in a fluid convection cell is constrained by the movement of all the other elements of the fluid (in a horizontal all-to-one sort of constraint). The whole of the system as such is more than the mere sum of its parts in the sense that the dynamics itself must also be accounted for by any faithful model of the system, but its causal power cannot be manifested over and above the causal power of each one of its parts if the trajectories in space-time that they follow are deterministic. Unless the behavior of the components obeys probabilistic laws that endow it with alternative futures, the dynamics of the system as a whole is only an emergent epiphenomenon and the interaction among particles is what really calls the shots.

---

[223] Sperry, R.W. (1980).

[224] "I hold that every time the elements of creation, whether atoms or concepts, are put together in the same way under the same conditions, that the same new properties would emerge and that the emergent process is, therefore, causal and deterministic" (*Idem*, p.200). The deterministic nature of the process of synchronic emergence is something that I am not questioning here.

[225] Popper, K., Eccles, J.C. (1977), p.35.

Let us look at a different diagram now:

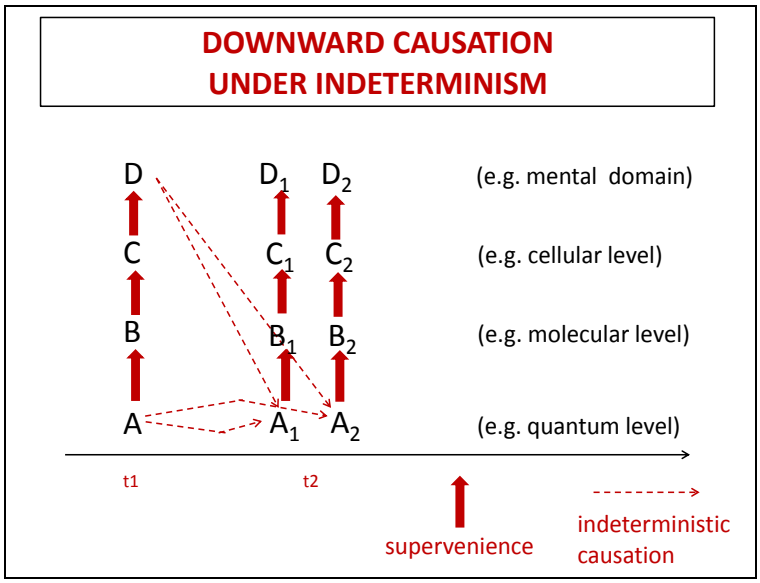


Fig.3: The upper-level states have downward causal power over the subvenient reality which is insufficiently determined by the previous physical states, given fundamental indeterminism.

In this diagram, the dotted arrows represent indeterministic causation, that is, a causal path that can either become actual or not. A_1 and A_2 are two alternative possible futures, caused by both same-level and top-down causes, each of which would have been insufficient to determine the outcome in isolation. Likewise, D_1 and D_2 are jointly caused by the hierarchical structure that synchronically leads from A_1 and A_2 all the way up to the highest level, and also, indirectly, by the D-level antecedents that made it so that the particular structure in question (either 1 or 2) was selected. So, even though supervenience holds, diachronic causal influences by the emergent entity described by state D have an effective power over which of the two alternatives (hierarchy 1 or hierarchy 2) will become actual. Edwards' dictum thus fails.

3.8. The break of supervenience and emergent indeterminism

The problem with my view, some might say, is that I am postulating a structural kind of relation between levels which presupposes natural supervenience (cf. 3.6.2). But if we

hypothesize that changes might happen at the chemical or biological level that are not matched by corresponding changes at the lower-levels, then there could be same-level causation at the chemical level, for instance, without the need for fundamental indeterminism. Actually, indeterminism would become an emergent feature in this scenario.

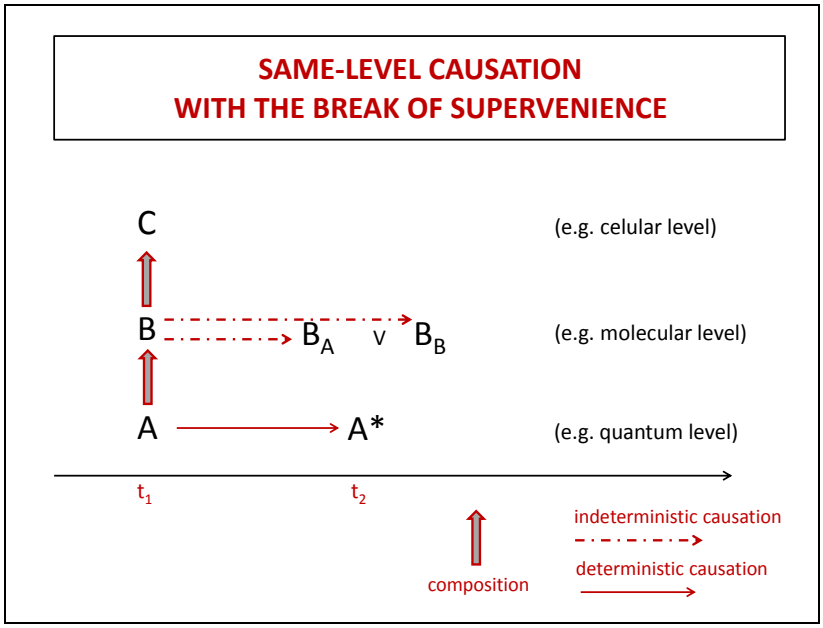


Fig.4: If supervenience breaks, then there can be a sort of "emergent indeterminism", for B would give rise to two alternative scenarios at t_2, even though there is a unique evolution at the fundamental level.

A problem with the above schema is that it lacks information in what concerns the relation between levels at t_2. The mereological relation that we could picture between A (e.g. physical particles and forces) and B (e.g. molecules at the chemical level, say) at t_1, is somehow broken at t_2, for it is hard to understand how A* can constitute both B_A (in the first alternative scenario) and B_B (in the second alternative scenario). It seems logical that a mereological sum cannot be identical both to X and to Y, if X and Y are different things. So the relation that could link A* to B_A or B_B would have to be an indeterministic relation, which leads us to the postulation of a causal relation (the only indeterministic relation we usually find in our metaphysics), as many emergentists today argue holds between an emergent and its emergent base[226]. Naturally, at t_1 the composition relation would have to be replaced with a causal relation as well, in our schema, for a matter of ontological

[226] Cf. O'Connor, T. and Wong, H.Y. (2005).

coherence. After these adjustments, this is how figure 4 could be completed and corrected:

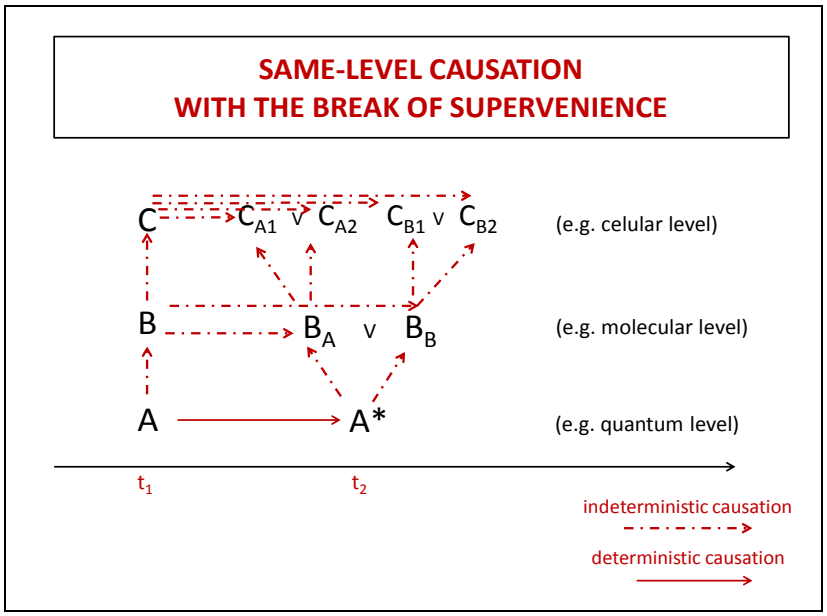


Fig.5: In a world were supervenience does not hold between levels, relations can be thought of as indeterministic causal relations, with branching possibilities from the bottom up.

Given the nature of indeterministic causation, at t_2 B could act as a partial cause of either B_A or B_B (in two alternative worlds), in collaboration with A*, which is also a partial cause of both. And the same could be thought to happen at the following levels in the hierarchy.

But is this a plausible picture of reality? Could indeterminism be emergent? Could there be some sort of "realizable multiplicity" – the opposite of "multiple realizability"?

A surprising aspect about this question is that many authors who deal with problems related to free will, indeterminism or emergence, do not really commit to either one of these two theses: a) that indeterminism can originate only from the bottom-most level, or b) that there can be some sort of "emergent indeterminism" in the world. This is very surprising, especially since this question is of the utmost importance for the free will problem – for if (b) is true, one could hypothesize the existence of alternative possibilities only at the mental level, with no need for any postulation about the nature of the physical substrate[227].

[227] This what Helen Steward argues should be done in her (2008).

In this hypothetical scenario of "realizable multiplicity", even if physics (in all its domains, classical or quantum) were strictly deterministic, the higher and more complex levels of reality that were disjunctively related to physical states could be diverse and the causal same-level sequences they would engage in could be described (if not governed) by probabilistic laws. Of course, these laws and the sequences of events at the upper domains that they would describe could not contradict the physical lower level laws. But if the latter were general enough and physics provided only necessary but not sufficient conditions for what happened in other domains, then unique evolution could be broken at a non-fundamental level.

Intuitively, one might tend to think this is something that we encounter on a daily basis, like in a soccer game where, apart from a small number of deterministic rules, there are many ways to play and score. And it is also what *prima facie* happens in the biological world, where the unfolding of events seems to have immense – if not infinite – possibilities, though constrained by a relatively small number of strict deterministic physical laws. In her influential 1971 essay on indeterministic causation, Elizabeth Anscombe also described this hypothesis elegantly, using the example of a game of chess: "the play is seldom determined, though nobody breaks the rules"[228].

However, if we try to build an analogy between these cases and the concrete and detailed examples we might find when analyzing a certain multilayered system like a cell or a molecule, we immediately see that we are comparing incommensurable things: coarse-grained descriptions of soccer or chess games and fine-grained descriptions of a physicochemical organic system such as a cell. The apparent looseness or generality of physical laws in the latter case is just a misunderstanding. When we move from the coarse-grained description of the macro-system to a more detailed fine-grained description that gradually includes all the relative values of all the material elements involved, then the "space of possibilities" gets increasingly reduced for the specific instant that we are considering, up to a point of full determination – unless we postulate

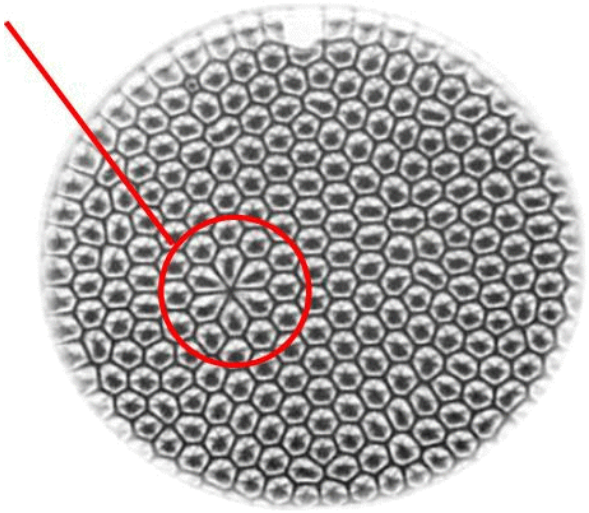
[228] Anscombe, G.E.M. (1971), p.99. Anscombe acknowledges that this comparison was first suggested by Gilbert Ryle [in *The Concept of Mind* (London, 1949), p.77], though his use of the example was different than hers.

probabilistic laws from the beginning. Certainly, such an inclusion of the micro values into the description of the system would make it cease to be the description of a *macro*-system. However, my point is precisely that if we want to assess if "the play" is "determined", we must go beyond the macro-descriptions and dig into the fine-grained laws that underlie the behavior of the players at each instant. The rules of chess or soccer are not the only set of laws at play in the field, just as the biological laws are not the only laws constraining the behavior of a cell.

As the reductionist research program pursued by most special sciences today is still in progress, we cannot yet be sure that only one macro-state can correspond to any micro-state of a system. However, that is the most plausible view, all things considered. Scientists, as well as ordinary educated people do not typically deal with reality as synchronically branching. If we put two glasses of milk in the refrigerator at the same time and one of them gets spoiled sooner than the other, we will first look for a visible difference between them: distinct packs with different expiring dates or some dirt in one of the glasses, for example. If we cannot find it, we will likely think there is an invisible reason that can explain what happened *and* explain also why it happened in only one of the glasses rather than in both. We will intuitively postulate a microscopic difference which will surely have caused the macro phenomenon that we can detect with our senses. In the same way, when scientists encounter differences between two samples of a similar substance at the biological level that they cannot easily explain, they will look for corresponding differences at the chemical level that might justify the phenomenon. Likewise, if a Rayleigh-Bénard cell shows a small local perturbation in the macroscopic expected pattern, that will be due to a small (maybe even microscopic) difference in the bottom surface of the container.

Fig. 6. Imperfection in a hexagonal convection cell pattern caused by a tiny dent in the plate.

[Van Dyke, M. (1982), p.83]



This is how science reasons and how it has progressed. Differences at one level lead researchers to look for differences at the next smaller one where explanations might be found, and along that trail they discover the patterns of the microscopic structure of reality, like the periodic table or the standard model of particle physics, and find out the genetic or viral origin of certain diseases.

To postulate that cases of emergence would break this principle and allow for upper-level changes to take place independently from micro-level changes would certainly fuel the case against the plausibility of emergence. And to appeal for the heterogeneity between physical causation and psychophysical causation in order to defend the break of supervenience in the latter cases only[229] seems too fragile a solution as well. That the mind depends on the brain and has a covariance relation with it has been the guiding assumption of neuroscientific research, a field in which extraordinary progresses have been made in the past decades. Today it is possible to infer from behavior and subjective reports the existence of chemical imbalances in the brain that can be treated pharmacologically, physical lesions that can be confirmed later via brain imaging

[229] Cf. O'Connor, T., Wong, H.Y. (2005).

techniques, etc. All this must be taken seriously into account when dealing with the associated philosophical problems.

Therefore, as a matter of prudence, I would avoid giving up supervenience. It would carry too heavy a cost, especially since I believe we do not have to do it in order to make sense of emergence. Assuming the truth of ontological indeterminism at the most fundamental level of reality is much more coherent with the current scientific picture of the world and thus comes surprisingly inexpensive as a solution for the dilemma introduced by Edward's dictum.

3.9. The irreducibility of the relation itself

One last thing is missing before we can move on to the next chapter. I have written several times that, in an emergence relation, the subvenient level "generates" or "produces" the supervenient one. But what does that production relation amount to? Is it a special sort of mereological relation, in which novel causal powers emerge but the bottom-level entities are parts of the upper-level one? Is it instead a dynamic, rather than static and formal, relation, in which the immediately preceding emergence base causally brings about the subsequent upper-level entities?

The first alternative would imply the assimilation of the emergence relation to the composition relations we find everywhere in nature. There is nothing fallacious in that, I believe, but it misses the point. To say that the emergent entity is composed by the elements that form its emergence base does not allow us to identify that which renders this relation different from cases where the systemic properties are simply resultant. In my opinion, a composite whole can be emergent as well as not, depending on whether its causal properties are explainable or not only on the basis of the causal properties of its parts. Likewise, as we said before, an emergent entity can be in a part-whole relation with its base, as well as not. In the cases where it is, the question of its "production" is given an easy answer: the emergent entity is brought about by the aggregation of the elements that compose it in a whole, together with a previously unmanifested and basic emergent

law. This seems to me to be logically possible – the question whether it is empirically plausible will be dealt with in the next chapter.

However, in cases such as the mind-body relationship, where mereology does not apply, how can the production or generation of the emergent entities out of the emergence base be understood? O'Connor and Wong[230] suggest that the emergence relation holding between the mental and the physical (which they assume to be fundamentally distinct) is a causal relation, albeit one that is not homogeneous to the physical-to-physical causal relation. They assume a dispositional account of causation according to which the appearance of a systemic emergent property is to be understood as the joint effect of the tendency of each of the basic entities to generate a collective effect that could never manifest below a certain threshold of organized complexity. It is thus unpredictable, but entirely natural.

The problem with this account is that, unless we wish to question the usual assumption that causes must precede their effects (which I do not want to do), it implies considering emergence as a diachronic dependency of the upper-level entity on its preceding lower-level cause. Emergent mental states, for example, would not be simultaneous with the neural states they depend on, but rather subsequent to them (even though the time lapse between them is supposed to be minimal). For example, the neural substrate of my mental state at t (say, a certain intention to act) would not be my brain state at t but rather my brain state at t_{-1}, while my brain state at t would be the neural substrate of my following mental state at t_{+1}.

This strikes me as odd. In a hypothetical snapshot of my complete self at t, there would be mismatch between my thoughts and feelings and my brain states. Imagine a temporal sequence in which my mental state M_A leads to mental state M_B , which leads to M_C; by hypothesis, at the neural level we would find N_B, then N_C, and finally N_D. In an alternative sequence (not necessarily an alternative world, but merely another moment in time in which, given different stimuli, say, a different mental sequence occurs), we can imagine

[230] Cf. O'Connor, T., Wong, H.Y. (2005).

M_X, rather than M_B, to follow M_A. At the neural level, in the first instant of the sequence, we would have N_X (the neural substrate of M_X) rather than N_B (the neural substrate of M_B). Now imagine M_A represents a question in someone's mind (e.g. "What am I to do?") and that M_B and M_X are two different answers to this question ("Do B" or "Do X"). In a hypothetical snapshot of my whole person in the instant when I am posing myself the question, I would find, in one circumstance, the neural substrate of the answer B, and on the other, the neural substrate of the answer X. But at the mental level, in both circumstances, I would still be undecided about the outcome.

My view instead is that non-mereological emergence is irreducible to other types of relation. It is a "building relation"[231] of its own, whereby a certain system belonging to one domain is necessarily linked to upper-level properties (or even substances) belonging to another domain.

A useful analogy can be that of color: at the strictly physical level no object possesses color, but once an object enters a relation with visible light and with an appropriate detector, such as the human eye (hence, given a certain context), it will necessarily manifest this secondary property. Color is not a new event or property *caused* by the primary properties of the object; it is *realized* by and *supervenient* on them, once the right trigger is in place. Thus, when I look at a red tomato, the tomato possesses both a certain superficial structure and the spectral reflectance that we call color, with the latter supervening on the former. Likewise, once the necessary and sufficient conditions are in place in a certain system (say, a mammal's developing brain), an emergent property appears (say, phenomenal consciousness), endowed with novel causal powers but nevertheless dependent on the lower-level structure in order to subsist.

The limits of this analogy are obvious: color is not an observer-independent instance of the ontology of the world. It depends on the interaction with a detector in order to manifest as something other than the primary properties of the object's superficial structure. Many even argue that, in fact, it is nothing but those primary properties[232]. This

[231] Again, I am borrowing this phrase from Bennett, K. (2011).

[232] Some of the discussion between physicalists, dispositionalists and subjectivists about color can be found here: Ross, P.W. (2000), Byrne, A. and Hilbert, D.R. (2004).

is not the place to get into that discussion. However, I believe the analogy can help us understand how different properties can be related by asymmetric dependency and simultaneous manifestation. When the *relata* at stake are truly distinct entities, we add to these modes of relation the underivability of the upper-level entity's causal powers and we are faced with ontological emergence.

In this chapter I have tried to show that the concept of ontological emergence is useful, compatible with our current science and that it can be given a coherent account. However, we still need to see if it corresponds to any actual features in the world. In the next chapter I will turn to this problem. Even though this is mainly an empirical question, I believe the appeal of the emergence hypothesis is greater in some fields rather than in others and thus its plausibility can be independently assessed.

4. THE CONSCIOUS SELF

> "How is it that anything so remarkable as a state of consciousness comes about as a result of irritating nervous tissue, is just as unaccountable as the appearance of the Djin, when Aladdin rubbed his lamp."
>
> (T.H. Huxley, 1866)

4.1. Emergence only at the conscious level

We have seen in the previous chapter that much evidence has been gathered that in nature, apparently, "more is different"[233]: as we go up the scale of complexity, our models have to include, at each level of organization, laws of a different type establishing relations between surprising new properties carrying causal powers that we cannot deduce nor explain only in lower level terms. Examples (though seldom uncontroversial) range from the liquidity and transparency of water, to physical phase transitions, to life and consciousness. Opinions differ a lot, however: some say emergence is ubiquitous, others consider it rare, others still inexistent.

For an emergentist like Robert Bishop, for instance, complex systems are permeated by constant inter-level relations by which upward determination is combined with downward constraint. Whenever a certain domain (say quantum physics) provides necessary but insufficient conditions for the description and, most importantly, the *existence* of phenomena at another domain (say, chemistry), contextual emergence is at work. These emergent domains or levels of organization, in which new constraints are superimposed over the broader and under-defined physical space of possibilities of the underlying levels are the subject matter of the special sciences.

[233] Cf. the above mentioned article by P. W. Anderson (1972).

> "In the absence of any other causes, physics supplies conditions defining the space of possibilities for matter's behavior and interactions. However, biological, psychological, and social realities further constrain this space of possibilities. These additional realities do not violate the space of physical possibility (i.e. never produce possibilities outside the physical space of possibilities). (…) For instance, the physical space of possibilities places relatively mild constraints on the motion of my arms, but my intentions in a voting context dictate when and how I will raise my arm in support of my favored candidate."[234]

According to Bishop, an example such as this one (the intentional motion of my arm in a voting situation) is no different from the example of clownfish, a hermaphroditic species in which the male located highest in the social hierarchy (the largest one) turns into a female when the dominant female of the group dies. The removal of the female from the social unit initiates a certain chemical mechanism (probably induced by the lack of pheromones) which is the triggering necessary condition for the next dominant male to undergo sex change; however, the size-based dominance hierarchy channels that trigger, providing one of the jointly sufficient conditions for which particular fish makes the switch[235]. In Bishop's view, this example from biology, along with examples from physics (temperature), chemistry (molecular shape) or the social sciences (how policies influence traffic, say), allows us to realize how the ubiquity of elementary physical laws does not imply the principle of the causal closure of the physical in its strong version, according to which such laws are sufficient to determine all behaviors.

I am skeptical about most of these examples, however. The contextual elements that seem to make all the difference can be interpreted in reductionistic terms, despite the greater simplicity of an upper-level explanation. In the clownfish example, the ecological reason why *this* specific male, rather than *that* one, changed into a female can most probably be translated into biochemical terms concerning certain reproduction-regulating hormones (even though the specific mechanisms that induce sex change in this species

[234] Bishop, R.C. (2010), p.607.

[235] Robert Bishop presented this example in the talk "Free Will and the Causal Closure of Physics" he gave in Lisbon (May 20, 2014), and discussed it with me in private correspondence.

are still under study). And what about all those examples of irreducibility in condensed matter physics that we analyzed in the preceding chapter? As I said, the singular limits that have prevented us so far from unifying physics under one universal theory of everything are indicative of the limitations of our models, but remain silent about the degree to which nature contains true cases of emergence or not. The only thing we know is that physical science (as well as the so called special sciences) is *compatible* with the conditions I presented as necessary for emergence to be possible:

- Physics cannot (and does not) prove that the microphysical world is causally closed
- Despite some controversy, the most common *interpretation* of quantum mechanics, our best theory about the most fundamental level of reality we know, is indeterministic, which means that in the quantum world, given a certain state of a system, different possible outcomes might follow.

So we certainly cannot exclude the possibility of emergence from the material realm; however, if we miss an independent reason to grant an emergentist explanation more credit than a reductionist one, we should probably bet on the latter. The success of science's reductionist research program in the last century justifies the confidence of most in the assumption that the full explanation of every fact in chemical-then-physical terms is just a matter of time. Every time we explore parts of the material world using the scientific method we find very complex and structured systems in which mereological pieces, like physical particles (and their properties), are combined into mereological wholes, like molecules and cells (and their properties). The living dynamics of a cell, for example, cannot take place at lower levels and it certainly appears as something new and unexpected, but it is explainable derivatively on the basis of the properties of the lower level components of the cell and their relations. Hence, even though our epistemic limitations and the discontinuity between our theories do not allow us to confirm nor refute reductionism, it is plausible that all material levels are resultant, as they stand in some kind of reductive relation to the fundamental domain. And for the sake of metaphysical simplicity, this seems to be the most reasonable position to assume whenever possible.

This reductionism-by-default position is very useful as it does not let us rest on the mystery. Instead of labeling the gaps in our knowledge as emergent phenomena, it rather forces into seeking deeply into the unknown, as in the famous Sidney Harris cartoon used by Dennett to taunt dualism:

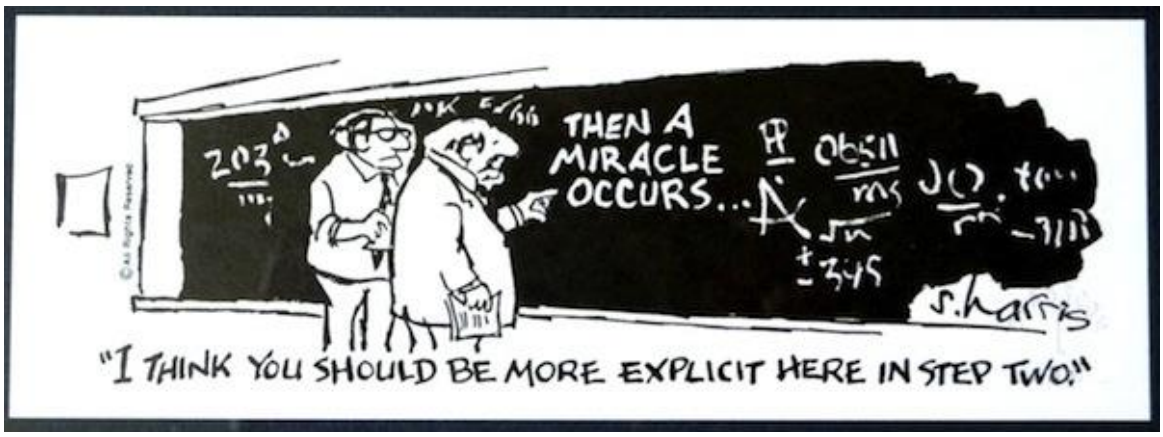


Fig. 7. ©1975 Sidney Harris – *American Scientist* magazine.[236]

But are there any cases in which we have an independent reason to consider that emergence is the only adequate concept we can use to characterize the relation between two levels of reality?

Phenomenal experiences are simultaneously linked to our bodies and radically distinct from them: they are dependent on the brain (if we drink too much alcohol, we immediately feel its typical effect on our conscious experience, for example), they are effective enough as to cause our body to move, but they appear to us by introspection as intrinsically qualitative and not amenable to a third-person description. Therefore, the relation between the body and phenomenal consciousness seems *prima facie* to be a promising terrain where to look for ontological irreducibility. Might we say that, in this case, one has more reasons to consider that there is a radical discontinuity in the very fabric of reality and not only in the eyes of the beholder?

[236] Cited in Dennett, D.C. (1993), p.38.

Even if the agent's physical body (including her brain) works as a machine the functioning of which can be exhausted by a reductive explanation, the agent's conscious states might be emergent. There is no reason to assume that if we "let emergence in" at a certain level of complexity, this should make us accept that there is emergence all the way down. Also, as action and free will are what most concerns us in this dissertation, what we need to assess is the possibility of a causally powerful entity that might choose and originate her actions without being determined to do so. And that does not require the falsifying of a reductionist account of material reality. What it does require is that the agent exerts her downward causal power over her physical body, and that her decisions might be caused by her, as an emergent substance, rather than by her parts. This relies only on the emergence of the conscious self.

However, one might ask why is it that the prudential reasons that led me to deny that epistemically emergent phenomena like living cells are ontologically emergent do not force me to embrace a reductionist view about consciousness as well. Indeed, a world in which a whole new type of reality (phenomenality) emerges at a certain point in evolution (and, synchronically, when a certain type and degree of complexity is reached) is certainly less parsimonious than a world in which everything, including our representations and feelings, is ultimately reducible to the most fundamental level. Defying Ockham's razor never comes in cheap. Nevertheless, conscious phenomena do seem to be of a totally different nature with respect to the physical world in spite of the fact that the latter provides the substrate from which they arise: they are subjective, qualitative and have a phenomenal feel to them that can only be directly experienced by the subject himself. In my view, arguments in favor of the emergence of phenomenality are much more powerful than arguments in favor of emergence in the non-mental realm. I will now turn to developing these arguments.

4.2. The irreducibility of consciousness

All the putative emergent phenomena that we considered in the preceding chapter appear in space, are quantifiable and can be detected by a third party, namely by machines. Conscious experience shares none of these properties.

Consciousness is the property of experiencing reality from a first-person perspective. It is the property we humans and, so we suppose, other complex animals are endowed with when we are awake and aware. There is something it is like to be *this* concrete conscious subject in this particular situation, and the what-it-is-likeness of this experience is radically subjective – I can know it only by having it. In the inner life we experience as conscious beings there is an epistemic asymmetry by which our conscious experiences are directly known to us and only indirectly translated into language and shared with others. There is no objective access to subjective phenomenal experience. As David Chalmers noted:

> "Even if we knew every last detail about the physics of the universe — the configuration, causation, and evolution among all the fields and particles in the spatiotemporal manifold — that information would not lead us to postulate the existence of conscious experience. My knowledge of consciousness, in the first instance, comes from my own case, not from any external observation. (…) Eliminativism about conscious experience is an unreasonable position *only* because of our own acquaintance with it. If it were not for this direct knowledge, consciousness could go the way of the vital spirit."[237]

None of the other properties we come in contact with in the world share this inherent subjectivity. We can know them by perceiving them, by measuring them, by calculating and deducing them, by learning about them from a third-person description. I do not need to experience being a dog in order to know how much my dog weighs, nor how his immune system works. And surely I can even infer from his behavior how he feels when I pet him. However, a future robot dog might look and behave exactly like my dog and miss

[237] Chalmers, D. (1996), pp.101-102.

consciousness altogether. Its behavior can never give us a complete (not to mention infallible) knowledge about its inner experience: I would need to *be* my dog in order to know *what it is like* to be my dog being petted. That aspect of my dog's reality is inaccessible to anyone besides himself.

Many arguments have been given in favor of this idea and of its implications. A very famous one is the one given by Thomas Nagel in his influential 1974 article "What is it like to be a bat?". Nagel argues that the phenomenal experience of being a bat, like the phenomenal experience of being human, is intrinsically subjective. It is impossible for humans to know how the experience of perceiving the world through a sonar system feels like, even if we know every bit of objective information there is to know about the mechanisms that make this type of perception work. The experience of echolocation is grounded on a particular point of view that only a bat or a very similar specimen can adopt and which is missing from any objective description. This is different from any other phenomena, even those of which we have qualitative experiences such as heat or color[238], because their objective character is independent from any particular point of view, and can be captured under a scientific description. Like Nagel says, "lightning has an objective character that is not exhausted by its visual appearance, and this can be investigated by a Martian without vision"[239]. On the contrary, phenomenal experience is, by its very nature, subjective.

The epistemic asymmetry by which conscious experience is directly accessible to the subject that experiences it and not to others led Nagel to infer the failure of the physicalistic ambition to reduce conscious experience to physical facts. The inherent subjectivity of conscious experience cannot be reduced to physical reality, which is objective by excellence. The gap is unsurmountable, for the translation of a subjective phenomenon into objective terms would annihilate the phenomenon in question.

[238] Of course, the experience of color is very often considered to be irreducible to its description in physicalistic terms [Cf. Jackson, F. (1986)], but we can easily distinguish the subjective experience from its objective source (the interaction of a certain receptor with the reflection of light with a certain frequency by the molecular structure of the surface of a certain material).

[239] Nagel, T (1974), p.443.

> "It is difficult to understand what could be meant by the *objective* character of an experience, apart from the particular point of view from which its subject apprehends it. After all, what would be left of what it is like to be a bat if one removed the viewpoint of the bat?"[240]

Nagel concludes that our inner experience of the world can never be satisfactorily accounted for via a third-person description.

However, some disagree and have endeavored to "explain" consciousness in purely physicalistic terms. Daniel Dennett, for example, in his book *Consciousness Explained*, purports to provide a functionalist model of conscious experience without adding any "mind stuff" to the objective reality that science can study – the brain. He acknowledges that "human consciousness is just about the last surviving mystery"[241], but believes that considering it to be something else besides the brain is to adopt dualism, and that this amounts to "just accepting defeat without admitting it"[242]. So Dennett undertakes a long journey through the mechanisms that lead to the construction of experience in which he tries to show that "there is no observer inside the brain"[243]. When we see and interact with the world, multiple drafts of information coming from our senses are registered in our memory and analyzed algorithmically (allowing us to identify edges, corners, faces, words), in the same manner as a small primitive robot like 1960's Shakey[244] can "perceive" reality, interpret it and react to it, without there being any conscious self inside it that "observes" some sort of "mental images" and has a first-person perspective about them. There is no projection of the items of phenomenology in a "Cartesian theater" somewhere in the brain. There is no audience for that theater, there is no appearance-reality distinction in human subjectivity, all there is are iterated discrimination processes that produce content that becomes available for eliciting behavior or for later memory retrieval.

---

[240] *Ibidem*.

[241] Dennett, D.C. (1993), p.21.

[242] *Idem*, p.41.

[243] *Idem*, p.110.

[244] Shakey was a robot developed at Stanford Research Institute by Nils Nilsson, Bertram Raphael and colleagues, whose mechanisms Dennett analyzes in (*Idem*, pp.85-95).

Dennett's reason for attempting to reduce phenomenality to objective reality is that he believes there is no other way of explaining it:

> "Only a theory that explained conscious events in terms of unconscious events could explain consciousness at all. If our model of how pain is a product of brain activity still has a box in it labeled 'pain', you haven't yet begun to explain what pain is."[245]

This would be a valuable project if it were possible; but I am afraid that, in the case of consciousness, to reduce the conscious *explanandum* to the unconscious *explanans* is actually just "a quagmire of evasion" and "petty word-jugglery"[246]. The result of our reduction would miss what is essential of the original concept, even if one uses the same word to name it. The reason is that phenomenality, the subjective and qualitative inner experience that a conscious person has of the world around her and within her, is something *other* than the functional aspects of our cognitive activity that can be somehow translated into the third-person quantitative and objective world of neurons and synapses, located in space and time and defined by structure and function. The phenomenal aspect of conscious experience may be produced by its neural substrate and be dependent on it to exist, but its causal reduction does not amount to an ontological reduction, because it is incommensurable with structural and functional concepts. This is why Chalmers distinguishes the easy problems from the hard problem of consciousness: the former can be given a functional explanation, the latter cannot.

> "It is an uncontested truth that we have the various functional capacities of access, control, report, and the like, and these phenomena pose uncontested explananda (phenomena in need of explanation) for a science of consciousness. But in addition, it seems to be a further truth that we are conscious, and this phenomenon seems to pose a further explanandum. It is this explanandum that raises the interesting problems of consciousness. To flatly deny the further truth, or to deny without argument that there is a hard problem of consciousness over and above the easy

[245] *Idem*, pp.454-455.

[246] These are two famous phrases that were used, respectively, by William James and Immanuel Kant, to describe compatibilism about free will.

> problems, would be to make a highly counterintuitive claim that begs the important questions."[247]

John Searle put forward an argument that can help us see this clearly. He starts by describing processes of successful reduction that have been made throughout history, regarding perceptual properties like heat, sound, color, solidity or liquidity.

> "In every case the ontological reduction was based on a prior causal reduction. We discovered that a surface feature of a phenomenon was caused by the behavior of the elements of an underlying microstructure. (…) In each case, for both the primary and secondary qualities, the point of the reduction was to carve off the surface features and *redefine* the original notion in terms of the causes that produce those surface features. (…) ''Real'' heat is now defined in terms of the kinetic energy of the molecular movements, and the subjective feel of heat that we get when we touch a hot object is now treated as just a subjective appearance caused by heat, as an *effect* of heat. It is no longer part of real heat. (…) If all subjective experiences disappeared from the world, real heat would still remain. (…) Part of the point of the reduction in the case of heat was to distinguish between the subjective appearance on the one hand and the underlying physical reality on the other."[248]

However, says Searle, the subjective appearances are the very essence of conscious experience. We cannot simply redefine it in terms only of the underlying physical causes, as in other cases, for the subjective and qualitative aspect of experience (qualia) is exactly what we are supposed to give a definition of, so it is not something that we can simply "carve off".

> "We can't make that sort of appearance-reality distinction for consciousness because consciousness consists in the appearances themselves. *Where appearance is concerned, we cannot make the appearance-reality distinction because the appearance is the reality.*"[249]

---

[247] Chalmers, D. (2003), pp.109-110.

[248] Searle, J.R. (1992), pp.118-121.

[249] *Idem*, pp.121-122. My reader might have noticed that both Searle and Dennett argue for the abandonment of the appearance-reality distinction when it comes to consciousness, but they go down opposite paths therefrom. Searle's recognition of the inadequacy of an appearance-reality

I have alluded before to Nagel's claim that there can be no "objective character" of a phenomenal experience:

> "It is impossible to exclude the phenomenological features of experience from a reduction in the same way that one excludes the phenomenal features of an ordinary substance from a physical or chemical reduction of it – namely, by explaining them as effects on the minds of human observers. (…) The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view."[250]

How can there be an observer-independent version of my mother's grief when her father died? Or of my feeling of overwhelming joy when my children were born? Of course, given our human ability for compassion and empathy, we can relate to other people's feelings and have an idea of what they are like, based on our personal experiences in similar situations. But our approximations to the feelings of others will never be *their* feelings. The qualitative and subjective aspect of their mental life cannot be reduced to the objective aspects that we can eventually reproduce and thus have a partial knowledge of. There can be an objective character of the content of a thought but not of the phenomenal representation of that thought in the mind of the one who has it. And certainly not of a sensation or a feeling. And since the first-person perspective is essential to phenomenal experience (it is *my mother*'s grief, *my* joy, *the bat*'s sensation of perceiving objects in the dark), such an experience cannot be identical to any portion of physical reality, which is by definition independent of anyone's point of view.

Pain is a good example of this impossibility. The definition of pain that is used both in medical and neuroscientific contexts around the world is the one provided by the International Association for the Study of Pain:

---

distinction leads him to acknowledging that the reality of consciousness *is* its appearance. In turn, Dennett's way of dropping that distinction is to reduce consciousness to objective "reality", eliminating the subject's inner life (the "appearance") from the model.

[250] Nagel, T. (1974), p.437.

> "An unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage."[251]

In a subsequent note, the authors of this definition state clearly, "*Pain is always subjective".* Thus, pain is defined in subjective and qualitative terms ("unpleasant experience"), even by scientists who strive to delimit it in an as objective and quantitative a way as possible. A functional definition of pain that would leave out its "what-it-is-likeness" could not possibly exhaust what pain is. It could account for the physical causes of the experience of pain, the behavioral consequences of it and the neural correlates without which pain is not possible. But it would miss the crucial element of how pain *feels*.

What happens in the case of pain, happens with any phenomenal experience: to analyze consciousness exclusively in functional terms is to collapse it with other concepts, such as, for instance, that of "awareness" – the functional notion of "a state wherein we have access to some information, and can use that information in the control of behavior"[252]. The concept of consciousness and that of awareness, however, are not logically coextensive. Even though they usually go hand in hand, there is nothing in the concept of awareness that implies the phenomenal aspect of there being something it feels like to be aware and thus I can conceive the possibility of being aware of a fact without having an associated phenomenal experience of it. Therefore, functionalizing consciousness would be merely "changing the subject"[253], not to mention assuming a highly counterintuitive position that brings with it the burden of proof.

That is why Dennett's *Consciousness Explained* can be easily nicknamed "Consciousness explained away" as the author himself, in the last chapter of the book, suggested could happen[254]. In the words of one of his many critics, his confessed enemy Searle:

> "To put it as clearly as I can: in his book, *Consciousness Explained*, Dennett denies the existence of consciousness. He continues to use the word, but he means something

[251] Merskey, H., Bogduk, N. (eds.) and IASP Task Force on Taxonomy (1994), p.209.

[252] Chalmers, D. (1996), p.28. What he calls awareness is what Ned Block labelled "access consciousness" (1995).

[253] Chalmers, D. (1996), p.106.

[254] Dennett, D. (1993), p.454.

> different by it. For him, it refers only to third-person phenomena, not to the first-person conscious feelings and experiences we all have. For Dennett there is no difference between us humans and complex zombies who lack any inner feelings, because we are all just complex zombies."[255]

Dennett is not troubled by this type of criticism. His view is that to explain things *is* to explain them away. We necessarily have to leave something of the *explanandum* out of the *explanans* if we are to explain it. "Leaving something out is not a feature of failed explanations, but of successful explanations"[256], he says.

However, deflationist positions such as his clash against what I believe to be an undeniable intuition expressed in Descartes' *cogito*. Even if I may be totally wrong about how the contents of my conscious experience relate to the external world (e.g. I am mistaken when I believe that the pain I feel is located in the my hurting limb rather than in my head), I cannot be wrong about the fact that I feel *something*. It is self-refuting to assert that one feels like a zombie. Thus, pretending to explain consciousness in physicalistic terms is actually denying the reality of the phenomenological experience one is supposedly trying to explain.

Reductive analyses of consciousness which leave out its subjective what-it-is-likeness fail to capture what consciousness is, also because they are "logically compatible with its absence"[257]. For instance, how can consciousness be identical with some neural state, if that neural state could logically be had by a zombie, a pseudo-person capable of objective reasoning, like a computer, but with no subjective experiences whatsoever? Thomas Nagel uses this argument in his 1974 article. Despite acknowledging that it is possible that, by nomological necessity, any device which is complex enough to create consciousness will in fact create it[258], he argues that this is not a logical necessity, hence it is perfectly conceivable that systems of functional states or intentional states "could be ascribed to

[255] John Searle, in his reply to Dennett's reply to his review of *Consciousness Explained* [New York Review of Books, "*The Mystery of Consciousness*: An Exchange" (December 21, 1995)].
[256] Dennett, D. (1993), p.454.
[257] Nagel, T. (1974), p.436.
[258] Cf. *Ibidem*, note 2.

robots or automata that behaved like people though they experienced nothing"[259]. This shows how functional or intentional states are compatible with the absence of consciousness and are thus not something it can be identified with or reduced to.

David Chalmers made a very systematic use of zombies in his case against physicalism[260]. The core structure of his argument is similar to Nagel's: according to Chalmers, we can conceive a being that is molecule-by-molecule identical to him and that behaves like him, but which misses phenomenal consciousness entirely. Even though he acknowledges that his zombie twin is most probably not physically possible in our world, its conceivability leads to the conclusion that it is metaphysically possible, which means that consciousness does not supervene logically on the physical: there could be a world that is physically identical to ours but in which there are no conscious beings. Or to return to Kripke's metaphor[261], after God created the physical world, he needed to work some more in order to add consciousness to it[262].

Last but not least, there is one more argument in favor of the irreducibility of consciousness to its physical substrate that deserves to be taken into account: Frank Jackson's "knowledge argument"[263]. Here, we are presented with Mary, a scientist confined from the day of her birth to a black-and-white room, where she learns all the physical (and chemical and neurophysiological) information there is to know about the world. One day, she leaves the black-and-white room and sees color for the first time. Jackson argues that her experience of seeing a red apple, for instance, is something new, something she had been missing, something which her complete knowledge of the physical facts about the world could not have given her before. Therefore, Jackson concludes, physicalism is false because qualia are not and cannot be included in a

[259] *Ibidem*.

[260] Cf. Chalmers, D. (1996, 2010).

[261] Kripke, S. (1972), pp.153-154. I mentioned Kripke's metaphor regarding logical supervenience in section 3.6.2.

[262] The zombie argument has been criticized both on the grounds that zombies are not conceivable [Dennett, D. (1995)], and that conceivability does not entail metaphysical possibility [Hill, C. S., and B. P. McLaughlin (1999)]. Chalmers has answered thoroughly to most of these objections, namely in his (1996) and his (2010).

[263] Jackson, F. (1982).

complete description of the world in entirely physical terms. Hence, if Mary's previous knowledge about vision and color (which, by hypothesis, was *all* the physical information to be had) was incomplete, then the conscious experiences that provided her with new knowledge cannot be satisfactorily accounted for by a neurophysiological description. They must be something over and above the causal interactions that take place inside our skull[264].

Dennett is much aware of this type of argument which argues for the ontological gap between the physical structure of the brain and conscious experience, based on the epistemic gap that relies on the radical heterogeneity between these two types of reality. And so he counters:

> "Why should consciousness be the only thing that can't be explained? Solids and liquids and gases can be explained in terms of things that aren't themselves solids or liquids or gases. Surely life can be explained in terms of things that aren't themselves alive – and the explanation doesn't leave living things lifeless."[265]

The analogy fails, however. Solids, liquids and living things can be explained in terms of things that lack their definitional properties because the *explananda* in such cases are phenomena that we wish to explain in terms of structure and function. Consider life, for example. Today, as well as in the heyday of vitalism, what we are looking for is the causal etiology of complex phenomena like adaptation, growth and reproduction, which may seem extraordinary with respect to our explanatory tools, but are not of a radically different nature. For instance, we can understand life as the name given to the homeostatic dynamics of an organism, and once we understand that dynamics in terms of structures, functions and laws of interaction, we might easily accept that there is nothing left to explain[266]. It is like explaining the whole in terms of its parts – a type of explanation that seems adequate and satisfactory because it exhausts the relevant

[264] As is well known, Jackson's argument led him to the conclusion that conscious states are epiphenomenal. However, since I do not assume the Causal Closure of the Physical to be true, I use the Knowledge Argument in favor of an anti-physicalistic thesis.

[265] Dennett, D.C. (1993), p.455.

[266] Many would disagree, of course. See for instance, Robert Arp's "Emergence in Biology" (2008).

features of the *explanandum*. As we have seen, with consciousness, things are very different. My feeling a certain experience is part of the essence of that experience, it is not dispensable as in the case of heat, say, which would remain present in the world had all the subjective experiences of heat disappeared.

Let us sum up. Any possible causal reduction whereby future science might describe the coming to be of conscious mental states may explain exhaustively the mechanisms of their production but will always fail to reduce them ontologically, for nothing else besides the conscious experience itself can contain what is essential about it.

Therefore, if to explain A is to restate it in terms of B, then we may as well admit that consciousness is unexplainable. As far as what our current conceptual framework allows us to see, no neuroscientific model, no matter how exhaustive, will ever be able to show us that conscious states are nothing over and above their neural correlates. Even if we identify their physical causes and understand in detail the conditions that have to be in place for them to emerge, conscious states can never be identical nor reducible to their non-conscious substrate.

This is why I believe that consciousness is the emergent property by excellence. The subjective, private and qualitative nature of conscious mental states cannot possibly be translated into third-person universal quantitative terms, such as the ones used to describe non-mental reality. So even if the self-explanatory character of emergent phenomena, with their mysterious novelty somehow transcending the causal processes underlying them, might seem too spooky to be included in any serious metaphysics, truth is that the reality of consciousness, which we are all intimately familiar with and cannot deny, possesses this exact same character: radical irreducibility.

4.3. The conscious self

But conscious states and events are not yet what we were looking for when we started investigating the emergence hypothesis. Consciousness is a property, not a substance,

and agent-causalism requires that the agent as a substance be irreducible to her parts, both at the neural as well as the mental level. But we can make a further step if we ask: *who* is it that possesses the property of being conscious? Common sense, as well as all the scientific evidence we have gathered so far, lead us into thinking that inanimate objects are not conscious[267]. Only living complex systems have this property. So what can the property of being conscious tell us about its peculiar bearers?

Some authors have argued that consciousness is a unified state that cannot be possessed by aggregates. Only the self[268] as an emergent unified substance can be endowed with it. I will now present this unity-of-consciousness argument in the words of William Hasker, one of its proponents, quoting him at length for the sake of clarity:

> "As an example of the unity of consciousness, I cite my awareness of my present visual field. This field includes the impressions from a set of shelves in the living room of my apartment, with books on the lower shelves and a number of plants (…). All this I observe without scanning or refocusing my eyes: momentarily, as it were. The visual field is not unified in any interesting aesthetic sense, but it is a fact that I experience it as a unity, all at once and not as a succession of discrete experiences. (…)
>
> Now, my procedure is to take a specific conscious state – the state I am in when I am aware of my visual field, as described above – and ask what physical entity it is that is in that state. (…) Let us say that it is my brain that is aware of my visual field, and I am aware of it in virtue of my brain's being aware of it. But not all of my brain need be involved. (…) Let V be the smallest part of my brain which contains the modeling of all the information from my visual field. (…)
>
> Should we say, then, that it is V which is aware of the visual field? (…) But if V is a whole composed of parts each of which is not aware of the visual field, how can V itself be aware of it? If we assume that each item of information is modeled in a

[267] Panpsychists, of course, would disagree.

[268] The self is a concept that is very hard to define, especially given the very diverse contexts in which the term is used (philosophy, psychology, psychiatry, neuroscience). I am taking it here as the self-reflexive subject of conscious states, the psychological core that is present in all the conscious states of a healthy person, who perceives the world from a certain perspective and refers to the owner of that perspective as "I".

> discrete subunit of the brain, we might suppose that each subunit is aware of the information it contains (…). [But] this does *not* enable us to explain the awareness of the entire field; this would be like saying that each student in a class knows the answer to one question on an examination, and that in virtue of this the entire class knows the material perfectly!
>
> (…) *A person's being aware of a complex fact cannot consist of parts of the person being aware of parts of the fact*. A conjunction of partial awarenesses does not add up to a total awareness."[269]

What Hasker is trying to show is that there is a property of the mind (conscious awareness of a visual field) which does not follow from the properties of the brain's parts (and their relations) which are thought to produce it. No sum of partial awarenesses can add up to total awareness. Analyzing the characteristics of a phenomenal experience such as the one quoted above reveals to us that there need be some entity that feels the simultaneous and unified "what-it-is-likeness" of the visual field. That entity is the self – whom I call the "agent" in the context of action production.

To say that there is a substance that we can call a self is not to say that the self is a concrete object or some supra-natural *thing*. What is crucial to the conscious self is unity. It can be a dynamical system that emerges from the interaction between its parts. Here is the difference between an aggregate and a system, according to Alicia Juarrero:

> "In an aggregate, the properties of the parts do not change depending on whether or not they are part of the aggregate. In a system, on the other hand, the properties of the components depend on the systemic context within which the components are located. (…) Correlation and coordination among the parts confer a peculiar unity on the overall system."[270]

Systems are not mere epistemic entities, they are present in our everyday reality, from cells and multicellular organisms whose identity is dynamically grounded on the interactions that keep them together and integrated, to the weather or the food we eat.

[269] Hasker, W. (1999), pp.125-128. Compare with a similar claim by Timothy O'Connor (2000, p.116).

[270] Juarrero, A. (1999), p.109.

These are all systems in which all the parts are correlated and depend on the relation they are embedded in, in order to be what they are. If we pick a flower from a tree, it will soon die; the flapping of the wings of a butterfly might influence a tornado thousands of miles away.

This is all familiar to us. The difference between these common non-emergent systems and possible emergent ones is the reducibility of the causal properties of the former, in contrast with the genuine novelty of those of the latter. So an emergent self does not have to be a hard core with no structure. It might as well be a dynamic system like many others in our body, whose unity derives from it being a system, but which is emergent because it can downwardly cause events in the material substrate that produces it in such a manner that one cannot take those causings as mere macro effects of many micro causings. Given the subjective and qualitative nature of conscious experience, the whole that emerges from the system's intrinsic and extrinsic relations is uniquely endowed with a first-person unified perspective on the world and with the ability to voluntarily control its own movements.

This happens with human as well as with non-human complex animals. Unlike the paramecium, which moves reactively only, animal agents are capable of controlling and directing their bodily movements and interactions with the world purposefully. And this faculty is based on the unified and subjective point of view with which their active self relates to the world and authors these actions. In the words of Helen Steward, who grounds agency in this body/owner distinction:

> "Something that can make its body do various things must be a thing of which it at least makes sense to say it 'has a body' – something that can reasonably be regarded as an 'owner' of its body. And it is only of some sorts of entity that it makes sense to say that they 'have' bodies, thereby separating what is moved (a body or a body part) from what is doing the moving (an animal). It is these entities that are potentially sufficiently complex to sustain an owner/body distinction which I call 'agents' (…). Which sorts of entity may be said to 'have' bodies (…)? Entities with a mind."[271]

[271] Steward, H. (2014), p.17.

Or I would rather say: entities with consciousness.

4.4. Is this dualism?

My naturalist readers will regard this thesis with dismay. To say that there are properties that are of a totally different nature from material properties is to adopt property dualism, which is suspicious enough; adding to this the idea that the self is an irreducible *individual* means to give in to full-blown dualism, today a largely discredited position among both philosophers and scientists.

True enough, I believe that the material and the phenomenal are two radically distinct pieces of reality, and so I embrace dualism in this sense. However, in contrast with Cartesian forms of substance dualism, I consider that the mind depends on the body, as it is produced by (rather than added to) it, and that it naturally supervenes on the brain in that there can be no mental change without a corresponding neural change. Therefore, on my account, the mind's autonomy is very limited.

One could name my form of dualism "non-cartesian substance dualism" after Jonathan Lowe, who defined this view as that according to which "persons or selves are distinct from their organic physical bodies and any parts of those bodies. It regards persons as 'substances' in their own right, but does not maintain that persons are necessarily separable from their bodies, in the sense of being capable of disembodied existence"[272].

Lowe argues that the self cannot be identical with its body as the identity-conditions of both are radically different. By this, he means that the logically necessary and sufficient conditions for the truth of any statement about the identity of a certain self are not the same as the logically necessary and sufficient conditions for the truth of any statement about the identity of the body it is associated with.

> "As evidence of this, it is very plausible to suppose, for example, that I could survive the gradual replacement of every cell in my body by inorganic parts of appropriate

[272] Lowe, E.J. (2006), p.5.

> kinds, so that I would end up possessing a wholly 'bionic' body, distinct in all of its parts from my existing biological body."[273]

He adds another consideration still: while it is clear that each one of my conscious mental states depends for its existence on *some* part of my brain (for if I were left headless for sure I would stop experiencing any sort of mental life altogether), it will become clear upon analysis that I am not identical with my brain nor with some part of it. In fact:

a) my brain *as a whole* is not the subject of *all* of my thoughts and other conscious mental states (if we remove some parts of my brain, there will be some conscious mental states that I can experience with the remaining parts)

b) there is no *part* of my brain on which *all* of my conscious mental states depend (as neuroscientific research has shown, brain/mental functions are specialized and localized)

On the contrary, I myself am the subject of all my thoughts and feelings, otherwise they would not be mine; I am what unifies them all in a continuum of personal experience. According to Lowe, this argument can be used as a *reduction ad absurdum* of physicalism, for the only way an exclusively materialistic account of the human person can answer it is by denying that there is such a thing as a subject of experience (like Hume did).

I agree with Lowe that the identification of the self with the animal body will always bump into this problem of personal identity and persistence over time. The most popular response to it is the criterion of psychological continuity, the idea that "you are, necessarily, that future being that in some sense inherits its mental features – personality, beliefs, memories, and so on—from you; and you are that past being whose mental features you have thus inherited"[274]. This is almost consensual. The controversy regards what these psychological features amount to and what they require. Could memory, intended as some sort of transplantable database, be sufficient to ensure identity? Not if that is the way we define memory. I believe that to treat the mental as something that zombies could have is to devoid it of one of its essential parts – phenomenality. I would

[273] *Idem*, p.9.

[274] Olson, E.T. (2007), p.17.

not be myself if my memory were transplanted into a computer with no feelings. All the data in my head might be there, all my past experiences, but not the what-it-is-likeness that accompanied them and which is what made them *my* experiences and not those of a hypothetical twin sister of mine who could have lived all these events along with me[275]. Therefore, the mental features which ground our psychological continuity and thus our identity are entangled with consciousness, a vision that goes back to John Locke:

> "Since consciousness always accompanies thinking, and it is that which makes everyone to be what he calls self, and thereby distinguishes himself from all other thinking things, in this alone consists *personal Identity*, i.e. the sameness of a rational Being: and as far as this consciousness can be extended backwards to any past Action or Thought, so far reaches the Identity of that *Person*; it is the same *self* now it was then; and it is by the same *self* with this present one that now reflects on it, that that Action was done."[276]

A whole dissertation would be needed in order to address this problem properly, so I cannot develop it any further. My intention in mentioning it is just to show how monistic accounts of the human person are more problematic than the physicalist orthodoxy would like to admit. Thus I certainly side with Lowe in considering that the phenomenal subject with whom I identify through introspection is distinct from the body that I own.

There is one aspect of Lowe's view, however, with which I disagree. The *non-cartesian* side of his account led him to considering that the relation between the self and its body is like that between a statue and the lump of clay out of which it is made (a very common example in discussions about building relations such as realization or constitution). The self is spatially extended:

[275] The possibility of transplanting not only the intentional aspects of my memory but also its phenomenal ones into a hypothetical computer endowed with the ability to feel is less clear. Could *its* new feelings be *my* transplanted feelings? This would imply that it could assume my point of view. But can a personal perspective be transplanted? I suspect this to be a contradiction in terms, but I can leave this question unanswered for now as it has no consequences for the point I am making about personal identity.

[276] Locke, J. (1689), p.335.

> "We ourselves, not just our bodies, occupy space and have properties of shape, size, mass and spatial location."[277]

According to Lowe, this is a common intuition that it would be important for any account of personal identity to preserve. I do not share this intuition, though. I do not feel that the body shape that my self is not identical to (according to Lowe's aforementioned argument concerning their distinct identity-conditions) should be definitory of my spatial features. Why should being six feet tall define me, if those characteristics are not part of my identity conditions? Sure enough, our current use of language makes us say that "I am six feet tall", not that "my body is six feet tall", but that does not mean that, upon reflection, one will not agree that the usual sentence is misguided. I also say that I am blonde, even though I know perfectly well that this feature is only accidental and that I might say tomorrow, after dyeing my hair, that I am now a brunette. The use of the verb "to be" in these sentences is very different from its use when we say, for instance "He is a mean man". In the former cases, it is actually referring to properties of our bodies, whereas in the latter it is referring to properties of ourselves. Thus *being* one's body features of shape, size, mass and spatial location is usually interpreted as merely possessing some accidental characteristics, whereas *being* some sort of person amounts to identifying oneself as the bearer of a certain psychological property which is enduring and not spatially extended.

In order to dissociate my view from Lowe's, I would rather choose another name for my account. The one that seems most adequate to me is "emergent dualism", the position defended also by William Hasker, whom I have cited above.

Hasker uses a suggestive analogy when he claims that the unified subject of conscious experiences is the emergent self:

> "A magnetic field, for example, is a real, existing, concrete entity, distinct from the magnet which produces it. (This is shown by the fact that the field normally occupies – and is detectable in – a region of space considerably larger than that occupied by the magnet.) The field is 'generated' by the magnet in virtue of the fact that the

[277] Lowe, E.J. (2008), p.9.

> magnet's material constituents are arranged in a certain way (…). But once generated, the field exerts a causality of its own, on the magnet itself as well as on other objects in the vicinity (…). *As a magnet generates its magnetic field, so the brain generates its field of consciousness."*[278]

This analogy should not be taken too far, of course. A magnetic field does not exhibit irreducible causal powers nor does it possess the kind of unity that the conscious self must have. However, the analogy helps us see how material entities of a certain kind can produce by some sort of emission[279] another type of entity, distinct from them despite its dependency.

Hasker tends to consider that, like a magnetic field, the field of consciousness too exists within a certain volume of space. He does not define it spatially on the basis of the properties of the body it is associated with, like Lowe did, but he suggests that the close natural connection between mind and brain makes it plausible that "consciousness is itself a spatial entity". Put in these terms, I have to disagree with him, on the basis of the reasons stated above. However, the dependency of the emergent self on the brain does put contingent physical constraints on the spatial area wherefrom it can exert its causal powers. In this sense, then, it is true that our emergent self is spatially confined and cannot exist as a supra-natural being.

To sum up, my account of the irreducible self who is the subject of our conscious mental states and the agent to whom actions can be attributed, has the form of a qualified dualism. It claims that the conscious self is distinct from its body, even though it is produced by it and depends on it for its persistence in time. The conscious self is an emergent individual whose activity supervenes on the brain while simultaneously determining its causal evolution, with downward causal powers.

[278] Hasker, W. (1999), p. 190. The quantum physicist Henry Margenau made a similar claim in his (1984): "The mind may be regarded as a field in the accepted physical sense of the term. But as a non-material field, its closest analogue is perhaps a probability field." (p. 97). I will address Margenau's influence on Eccles' dualist account of the mind-brain relation in section 4.6.2.1.

[279] Cf. Geral Vision's "emission" account of emergence (2011, p.47).

4.5. Neural indeterminism

My account of ontological emergence requires that there be genuine indeterminacy at the level where the causal power of the emergent entity is supposed to be effective. In the case of an agent, this means that the human brain must work indeterministically for mental causation to be possible.

Many might object to my account with the claim that, at the level of neural events, there is no room for genuine indeterminacy, even if contemporary physics accepts it exists at the level of particle interactions. Quantum effects are washed out when it comes to large numbers and the brain is a machine that no scientific evidence has ever shown works indeterministically.

I will give a twofold answer to this objection. First, I will argue that, even if we cannot know whether our neural activity leaves room for alternative possibilities, there is no scientific evidence to this day that the human brain works deterministically either. Second, I will present scientific evidence that quantum events can have macroscopic effects.

What neuroscientists deal with on a daily basis are stochastic processes. The question whether these processes are only indeterministic at an epistemic level or are actually the macro manifestation of more fundamental indeterminacies is something empirical research cannot tell us just as yet. Adina Roskies, philosopher and neuroscientist, explains this very clearly:

> "The picture that neuroscience has yielded so far is one of mechanisms infused with indeterministic or stochastic (random or probabilistic) processes. Whether or not a neuron will fire, what pattern of action potentials it generates, or how many synaptic vesicles are released have all been characterized as stochastic phenomena in our current best models. However, whether the unpredictability we perceive is really due to fundamentally indeterministic processes, or to complex deterministic

> ones beyond our present understanding is something neuroscience cannot tell us."[280]

The problem with expecting science to give an answer to the question of determinism is that it is actually a metaphysical question, not an empirical one. Our quest for the origin of phenomena deep into ever more fundamental layers of reality is never concluded: one can always postulate one more level underlying the ones we have come to know well. That is why Roskies says that neuroscience will *never* be able to give us a definitive answer to the mystery of neuronal indeterminism:

> "Because a deterministic system can radically diverge in its behavior depending on infinitesimal changes in initial conditions, no evidence for indeterminism at the level of neurons or regions of activation will have any bearing on the fundamental question of whether or not the universe is deterministic. That is ultimately a question for physical theory, and will be answered by our best theory of the fundamental nature of physics, not at the level of brain science."[281]

So even if neuroscientists deal with epistemic indeterminism at the neural level, the nature of the causal interactions taking place underneath are out of their reach. And maybe the causal nature of the world will always remain ultimately inaccessible to the observer, even at the microphysical level of analysis.

Nevertheless, the fact that what we know so far meets exactly what would be expected *if* the brain worked indeterministically deserves to be acknowledged. Neuroscience has not discredited the hypothesis of neurological indeterminism and it keeps accumulating evidence that is entirely consistent with it.

Moreover, the fact that this question has not been settled yet has not prevented major neuroscientists such as William Newsome[282] or Paul Glimcher[283], along with many

[280] Roskies, A. (2006), p.420.
[281] *Idem*, pp.420-421.
[282] Newsome, W. (2014).
[283] Glimcher, P. (2005).

others[284], from expressing their conviction in favor of an indeterministic account of the nature of brain processes. From their point of view, the indeterminacy that we find at the behavioral level is a result of genuinely random events at the cellular and subcellular levels, like for instance the patterns of vesicular release and the variations in membrane voltage, which seem to be "the product of interactions at the atomic level, many of which are governed by quantum physics and thus are truly indeterminate events"[285]. In the article "Indeterminacy in brain and behavior", in which he presents a long review of the neuroscientific literature related to the problem of the source of variability in the brain, Glimcher concludes:

> "Physical indeterminacy seems to be a fundamental property of the brain."[286]

Of course, this depends on the general possibility of quantum fluctuations having macroscopic effects in warm and wet environments such as the brain, which is the second aspect I committed myself to providing some evidence for at the outset of this section. Some years ago this might have seemed much harder. Phenomena like superconductivity or the Bose-Einstein condensates were usually cited as examples of indeterministic effects at the macro scale but they required temperatures close to absolute zero. Today, however, there is increasing evidence that functional quantum effects operate in biology as well[287]. I will now provide two strong examples that have been in the spotlight lately.

The first case regards electronic quantum effects which occur at ambient temperatures in proteins involved in photosynthesis[288]:

---

[284] Most notably, Björn Brembs has been focusing part of his neuroscientific research on the intrinsic source of behavioral variability in biological organisms, a topic on which he has a clear position in favor of neural sensitive dependence on quantum fluctuations [Cf. Brembs, B. (2011)].

[285] Glimcher, P. (2005), p.49.

[286] *Ibidem*.

[287] In a recent article (2014), Stuart Hameroff and Nobel laureate Roger Penrose, probably the two most famous defenders of the idea that consciousness is a product of quantum effects in the brain, cited many studies which purport to show evidence of quantum effects in biological processes such as: ion channels, sense of smell, DNA, protein folding, and biological water (p.63).

[288] Engel, G. S. *et al.* (2007), Lee, H. *et al.* (2007), Mercer, I. P. *et al.* (2009), Collini, E., *et al.* (2010), Hildner, R. *et al.* (2013).

> "Light-absorbing molecules in some photosynthetic proteins capture and transfer energy according to quantum-mechanical probability laws instead of classical laws at temperatures up to 180 K. This contrasts with the long-held view that long-range quantum coherence between molecules cannot be sustained in complex biological systems, even at low temperatures."[289]

The second example of warm quantum effects was discovered in bird brain navigation, where "the ability of migratory birds to orient relative to the Earth's magnetic field is believed to involve a coherent superposition of two spin states of a radical electron pair"[290] (a spatially-separated pair of correlated electron spins). The mechanism underlying this phenomenon is currently under debate[291], but it is quite uncontroversial that we are before biological processes in which the effects of quantum phenomena are amplified and have chemical consequences[292]. The greatest novelty of these findings is the fact that, in these living systems, quantum superposition and entanglement are sustained for at least tens of micro-seconds – a much longer duration than that found in man-made molecules. This is a necessary condition for quantum interactions to have effects at the macro-scale[293].

Even though these are empirical studies which are clearly out of the range of the present discussion, they show us how the scientifically founded objections to an indeterministic brain are much weaker now than some years ago. They rely on the idea that micro-scale indeterminism would be cancelled out in a warm and wet system like the brain, but there is abundant evidence in favor of the presence and efficacy of quantum effects in other biological systems, which shows how sensitive dependence to quantum fluctuations in the brain is actually plausible.

---

[289] Collini, E., *et al.*, p.644.

[290] Walters, Z.B. (2014), p.1.

[291] E. Gauger et al. (2011), Bandyopadhyay, J. N. (2012), Walters, Z.B. (2014).

[292] There is in fact a new field of research called Quantum Biology, dedicated to studying non-trivial quantum phenomena in biological systems [Cf. Huelga**,** S.F.**,** Plenio**,** M.B. (2013)].

[293] Max Tegmark, for example, suggested that any macroscopic quantum entanglement in the brain would be destroyed in times of the order of 10^{-13} to 10^{-20} seconds. [Hodgson, D. (2012), p.145].

The concept of plausibility I am mostly interested in is naturalistic plausibility. According to Christopher Franklin, plausibility has to do with the demandingness of a theory's commitments:

> "[In order to assess a theory's naturalistic plausibility,] in addition to considering the *quantity* of empirical commitments that a theory has, we must also consider the *quality* of such commitments — specifically how demanding they are. The demandingness of a commitment is partly a function of how radically things must change from what we currently take ourselves to know in order for the commitment to be satisfied."[294]

What I hope to have shown in this section, and to keep providing evidence for in the following one, is that the empirical commitments of an agent-causal theory of action are much less demanding than is usually assumed. What mainstream science currently considers to be true would not have to change in order to accommodate the agent-causalist's requirements. It will only have to increase the resolution of the scientific picture of the world that is on the table, adding to it the further details that future research is expected to bring.

4.6. Downward causation from consciousness to brain

Despite the optimistic conclusions of the last section, the million dollar empirical question remains as to *how* can the conscious self exercise her downward causal power over the brain. Is this some sort of magic?

One can imagine this influence being exercised by the conscious self on any one of the underlying levels, or directly on the bottom-most level. As I explained in the last chapter, we have no way of knowing whether the epistemological emergence we find between strictly material levels corresponds to an ontological emergence. Also, for the reasons I presented in the first section of this chapter, I believe the best attitude to adopt in what

[294] Franklin, C. (2013a), p.128.

regards the material world is what I called reductionism-by-default. This, however, bears as a consequence that the levels of organization of reality are regarded as reducible to one another and thus synchronically determined from the bottom up, which entails that we cannot postulate a direct downward causal effect of the mental on the biological, for example. The mental has to have control over the bottom levels on which the biological level supervenes.

Therefore, I believe we should opt for a model which endows the conscious self with a power to affect directly the bottom-most domain (the quantum level, as far as we can tell). This means that, when there are alternative possibilities at the quantum level, the conscious self must somehow intervene to determine which of these alternatives becomes actual. The following diagram will make this clearer:

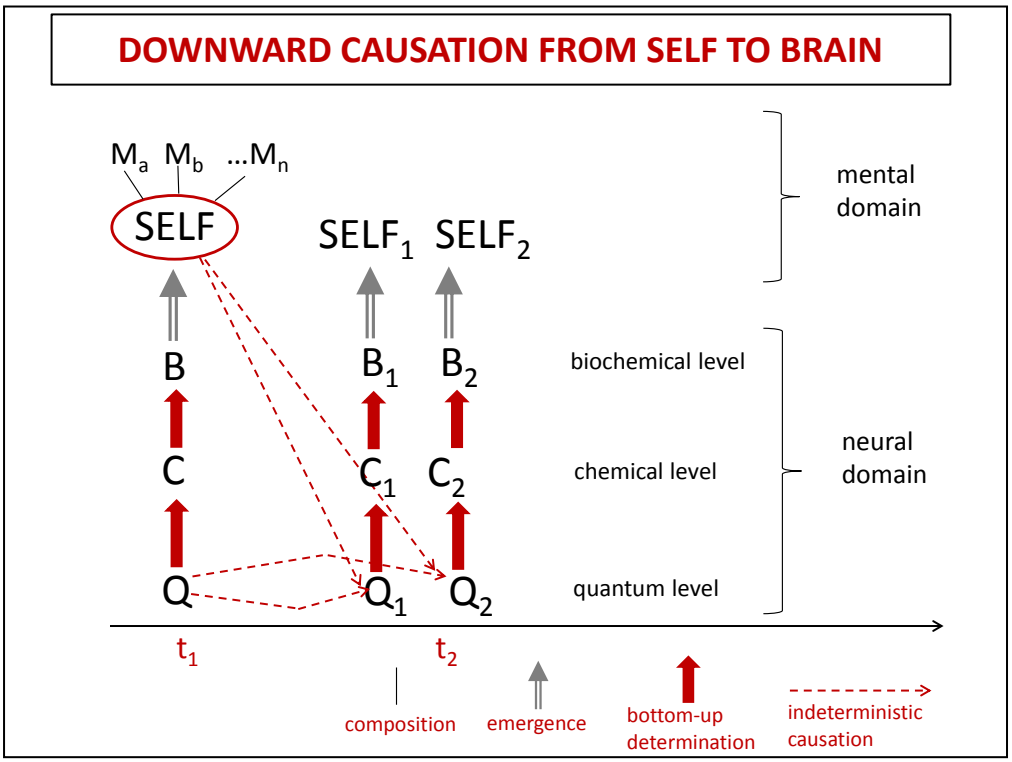


Fig.8: Synchronically, the conscious self (composed of different mental states M_1-M_n but endowed with a new causal power that none of them individually has) emerges out of the neural multi-layered substrate. The self is an effective cause that can determine at t_1 the actualization at t_2 of one of the different quantum states that at t_1 have a certain probability of being partially caused by Q at t_2. Depending on the intervention of the conscious self, either the 1st or the 2nd psychophysical reality will happen.

In this schema we can see how, even if the material world is structurally organized in such a way that each hierarchical level fully determines the next one in an upward manner, there is room for the emergence of the mental domain where a substantial self simultaneously contains and transcends her diverse mental states. The conscious self is a unitary individual brought about by the complex organization of the physical brain. The fact that the physical world is not causally closed and that quantum laws are only probabilistic allows it to exercise its emergent causal powers over the material reality, contributing at a given instant t_1 to the production of a new state of affairs at t_2. In addition to these conditions, which we assessed in chapter two, this view of psychophysical causation relies on two more assumptions: a coherent metaphysical account of substance-causation and the empirical possibility of quantum events being influenced by non-physical entities.

I will now assess them both.

4.6.1. Substance-causation

Causation is one of the most intractable problems in philosophy. Many different accounts of causation have been given in the years and there is no prospect of consensus[295]. As philosophers very well know, at least since Hume, causality is opaque. It is not a transparent relation between objects, properties or events that one might be able to experience with the senses or to somehow infer *a priori*. If we try to analyze the following causal process – a bridge collapsed after the explosion of a bomb[296] – we will find that it can fit many different descriptions. But which one is more faithful to the underlying reality? What caused what, in this series of events? Did the bomb cause the collapse of the bridge? Did each of the bomb's particles cause each of the single events that collectively amounted to the collapse of the bridge? Did the event of the explosion do the

[295] Cf. Beebee, H., Hitchcock, C., Menzies, P., eds. (2009).

[296] This example is used in Lowe E.J. (2008), p.3, 122.

causing? Did the single events (e.g. each atom's energy release) do it? While all these sentences seem to be suitable explanations for what happened, the metaphysical problem of detecting which one of them is more fundamental (so that all the other descriptions can be reduced to it) is not simple.

The standard view of the causal relation takes its *relata* to be events[297] or states of affairs[298]. One of the problems with the sort of agent-causation depicted in Figure 8, then, is that it seems to imply that in cases of action production an exceptional form of causation (that in which the cause is a substance) takes place, which renders these cases intrinsically different from all other causal processes in nature. While some accounts[299] have no trouble assuming this heterogeneity, such a discontinuous view has been accused of implausibility. And in fact, an account of causation will definitely be more elegant and parsimonious if it manages to unify several phenomena under the same principle. The different horns of this objection have probably led recent agent-causal libertarians to either assume that all causation is ultimately substance-causation[300], or to adopt a pluralist account of causation, according to which there are diverse categories of cause in the natural world (events, facts and substances), in each of which case causation is to be understood differently[301].

Views on this matter usually depend heavily on one's preferred ontology, and my case is no different. I favor an ontology according to which the world we live in is made of *substances* (persisting entities whose identity does not depend on anything other than themselves[302]), which are the bearers of *properties*. It is these properties (such as electric charge, mass, color, liquidity or phenomenality) that carry the causal powers that each substance has and, in doing so, direct the substance towards certain effects. Substances are essentially active because the properties that make them what they intrinsically are

[297] Cf. Davidson, D. (1980), Lewis, D. (1986).

[298] Cf. Armstrong, D. (1997).

[299] Cf. O'Connor's view developed in *Persons and Causes* (2000).

[300] Cf. Lowe, E.J. (2008).

[301] Cf. Steward, H. (2014).

[302] Cf. E.J. Lowe's definition of substance (2006, p.5). Note that substances can be tiny fundamental simples as well as more complex emergent compounds.

endow them with certain dispositions and liabilities by which they can affect and be affected by other substances around them, and thus make a difference in the world[303]. Causation consists in this difference-making and the laws of nature are the description of these regular interactions between substances, given their powers. Laws are not what *pushes* things into behaving in a certain manner. Things are internally directed towards their most probable behavior.

Powers cannot be exercised without a substance that instantiates them and, vice-versa, a substance cannot exist apart from the properties that identify it. So we can say that what causes effects in the world are substance-power pairs.

A concern that one might have regarding a substance-causal view of this sort is that it might be only the product of ambiguity:

> "It is not unnatural to think of substances generally as causes, but such thoughts, more carefully stated, are about events involving those substances, or states or property instances of those substances, as causes. (…) The concern is that we may not have a conception of substance causation that cannot be reformulated as event-causation. If this is in fact so, then it appears that the agent-causalist is providing only an empty verbal solution."[304]

However, there is no emptiness here. There is a perfectly clear sense in which a ball, because of its solidity and weight, has the capacity to smash a window, under the right circumstances (by hitting it with a certain speed). Given its mass and density, my body has the power to squash the grass beneath my feet, pressing it with a certain force which causes my footprints to be marked on the ground. An electron has the disposition to repel another electron. There is no language manipulation in saying that when the repelling event does happen, it consisted in the first electron's exercising its power to repel. This does not mean that these propositions cannot be "reformulated" in event-causal terms. I think they can, I just cannot see what good can come out of it.

[303] My view is inspired by Jacobs, J.D., O'Connor, T. (2012).

[304] Pereboom, D. (2004), p.11 of the manuscript. In the cited passage, Pereboom is presenting but not endorsing this objection, which was put forward by Alfred Mele (2006).

It seems to me that to analyze these cases in event-causal terms would be to describe the sequence of events from a certain point of view that we can call *external* – "My body's pressing the grass caused its squashing" – which can be explanatorily useful, but is actually just elliptical for instances of regular substance-causation. In contrast, the *internal* point of view that looks at the causal process from the perspective of a substance and its powers is the one that reveals to us the forces that drive causation in the world. Events cannot cause anything because they consist in the temporal succession of what happens, not in the subjects involved. Only the objects in the world have the right type of properties for causing certain events. Only the bomb is appropriately explosive, and it is that property that causes the explosion of the bridge (this causing itself being an event that can be described).

Another objection to substance-causation, however, seems to be more disturbing. It stems from the following argument by C. D. Broad:

> "It is surely quite evident that, if the beginning of a certain process at a certain time is determined at all, its total cause *must* contain as an essential factor another event or process which *enters into* the moment from which the determined event or process issues. (…) How could an event possibly be determined to happen at a certain date if its total cause contained no factor to which the notion of date has any application? And how can the notion of date have any application to anything that is not an event?"[305]

There are two ideas underlying this objection: first, that insofar as an event is caused to happen at a certain time, the cause that determined the timing must itself be dated; second, that a dated cause must be an event, because only events are dated entities.

I suspect that both these assumptions can be called into question by sophisticated arguments but I do not intend to do so myself. I actually consider them to be quite plausible. What is crucial for my argument is the notion of "total cause". Broad is assuming

[305] Broad, C.D. (1952), p.215. Carl Ginet stated a similar objection in terms of explanation in his (1990, pp.13-14).

that substance-causation works in a context-free manner, excluding any reference to time. But this is not the only way one has to conceive of the substance-causal process.

On my view, the possibility of a substance's exercising its causal power is structured by the circumstances, which is to say it depends on events that take place in a certain time and place. Sugar is always soluble, but it only dissolves if and when it is immersed in a liquid. That dated circumstance is what allows for that property, and the causal power it carries, to produce the effects it does on the involving substances (e.g. the particles of the liquid).

Let me use again the example of the explosion of the bridge. The bomb is what caused the explosion[306] due to its properties. However, that causing only took place when the circumstances allowed for it – e.g. when the right dose of pressure on the bomb triggered the manifestation of the disposition it had to blow up. The fact that a certain event (the exertion of a certain pressure on the device) had to happen for the causal power of the bomb to be released does not entail that it was that event that caused the explosion. Only that substance-causation is a contextual phenomenon and contexts are given in time, in the succession of events by which things change. The "total cause", we may say, includes that context.

The same applies to agent-causation, a complex species of substance-causation. The conscious self's power to affect the unfolding of physical events is what we perceive as the agent's power to decide and to act in a certain way rather than another. This power is partially determined by the agent's standing values, preferences and traits of character, together with her present reasons and other motivations (the mental states depicted as M_1 to M_n in fig.8). All these elements endow the agent's self with certain dispositions, more or less deterministic, depending on the degree of uncertainty in the choice to be made. However, being an emergent entity, the agent's self amounts to more than her reasons, preferences and so forth: her causal powers are not merely derivative. And since the powers associated with the agent's properties are not unconditional, they depend on

[306] Whether what causes the explosion is each single particle or the bomb as a whole depends on whether the causal powers of the bomb are resultant or emergent, as I explained in chapter 3.

certain circumstances, like the solubility of sugar manifests only when it is immersed in liquid.

When the agent's propensities for acting in a certain way are not deterministic, there are, by definition, no sufficient reasons for a certain action or decision[307]. In fact, the problem of understanding *when* something takes place is common to many instances of indeterministic causation, independently from the account of causation one assumes. Think of radioactive decay, for instance, in which the emission of radiation by an unstable nucleus can happen at any time. After the moment when all the necessary conditions are in place (which, in the case of agent-causation, means that the reasons that structure the agent's propensities have been acquired and are actively conditioning her decision-making), the substance at stake has a certain propensity to cause a certain effect. There is a certain probability that the causing will happen rather than not, that it will happen now rather than latter and that the substance-cause will bring about *this* effect rather than *that*. But there may not be a contrastive reason that explains any of these alternatives[308]. And that, I repeat, is the mystery of indeterministic causation in general.

To sum up: an ontology of substances and powers enables us to regard substance-causation as the regular and universal form of causation in the world. Downward causation from the mental to the physical, then, is just a particular subset of causal events in which the *relata* belong to different domains and have radically different natures. This might make us question the plausibility of consciousness having effective causal powers over material entities, which is what will be assessed in the next section, but it is not

[307] In her article "Causality and Determination" (1971), Elizabeth Anscombe took advantage of the ambiguity of the notion of sufficiency in ordinary language in order to make this clearer: «'Sufficient condition' is so used that if the sufficient conditions for X are there, X occurs. But at the same time, (…) 'sufficient condition' sounds like: 'enough'. And one certainly *can* ask: 'May there not be *enough* to have made something happen - and yet it not have happened?'"» [Anscombe, G.E.M. (1971), pp.90-91]. Anscombe notes that *sufficient conditions* is a term of art, and that we do not have to interpret it as *necessitating conditions*. However, it seems to me that what the author does is to replace "necessitating" with "necessary". In other words, by suggesting that *sufficient* could be intended as *enough*, Anscombe is referring the concept to the conditions that need to be in place for the effect to be possible (necessary conditions), instead of the conditions the presence of which forces the effect to happen (sufficient conditions).

[308] About the problem of the lack of contrastive explanation in cases of indeterministic causation in general and of undetermined actions in particular, see section 5.3.2.1.

substance-causation as such what distinguishes these causal interactions from other, strictly physical, ones.

4.6.2. Consciousness and the quantum world

Now we have to address the very concrete question of how plausible it can be that the conscious self can interact with physical reality at the quantum level. Expressed in different terms, it is the same question princess Elisabeth asked Descartes:

> "Given that the soul of a human being is only a thinking substance, how can it affect the bodily spirits, in order to bring about voluntary actions?"[309]

Given their material monistic assumption that mental states are identical to physical states, together with a reductionist account of the self, the majority of scientists believe that mental causation is nothing but physical causation. So the problem princess Elisabeth raised is nothing they should care about: there is no mind-matter interaction intended as the interaction between two distinct entities. Nevertheless, there are some exceptions among eminent scientists, such as Roger Penrose and Stuart Hameroff[310], Henry Stapp[311], and several other scientists who have dedicated their work to exploring the hypothesis that consciousness is an essential and irreducible part of the natural world, and who consider that quantum physics is the adequate framework for understanding the emergence and activity of this property. According to them, quantum theory can provide us with the ultimate answer to the age-old question of the mysterious interplay between the "matterlike and mindlike parts of nature"[312].

Empirical work is not the object of this dissertation but I believe there are two scientific hypotheses that were foundational to all subsequent work in this area and which deserve a more extended reference: one by John Eccles and one by John Von Neumann and

[309] Princess Elisabeth of Bohemia's letter of May the 6th, 1643.
[310] Penrose, R. (1989), Hameroff, S., Penrose, R. (2014).
[311] Stapp, H.P. (1993, 2006).
[312] Stapp, H.P. (1993), p.vii.

Eugene Wigner. Their proposals allow us to understand two crucial facts for the mind-brain interaction:

1) That there are micro processes in the brain which we can describe only probabilistically and which are simultaneously small enough for quantum influences to be significant, and embedded in networks that amplify those effects;
2) That quantum mechanics is an epistemological theory which provides a description of a system before and after a measurement but which cannot explain the transition between the probabilities before, and the defined state of the system after. That is where different interpretations come in, and some of them consider consciousness to be a fundamental element in this process.

4.6.2.1. John Eccles' dualist interactionism

The first scientific account I want to present was put forward by Sir John Eccles, a successful neurophysiologist who dedicated his whole life's work to understanding the mechanisms whereby the immaterial self can control its brain[313]. His thorough knowledge of the functioning of the brain, especially of the processes of synaptic transmission (for which he was awarded with the Nobel Prize in 1963), enabled him to develop a comprehensive and detailed theory of the mind-body relation, in which he strived to render his dualist convictions consonant with brain science.

According to Eccles, mind and brain are two independent entities that interact by means of quantum physics. Their interaction is enabled by the existence of genuinely indeterministic neural processes in the brain, crucial to its functioning, the probabilities of which could be influenced by the self. Eccles' life-long work was dedicated to better understanding the mechanisms underlying this influence.

Eccles centered his attention on exocytosis, the process in which a vesicle containing neurotransmitters is released into the synaptic cleft (the space separating two

[313] I have already alluded to his work together with Popper in section 3.7.

communicating neurons). This momentary opening of a channel in the membrane of a bouton (the terminal of the presynaptic neuron) is caused by a nerve impulse that results in a large influx of Ca^{2+}. Given the input of four Ca^{2+} ions to a synaptic vesicle, exocytosis may occur with a certain probability.

Exocytosis is the "basic unitary activity of the cerebral cortex"[314]. It causes neurotransmitters to be released, which then bind to receptors on the postsynaptic cell. The opening of the channels gated by those receptors causes transient changes in the membrane potential, the summation of tens or even hundreds of which is required for an action potential to occur (the discharge of an electric impulse). This spike will then travel along the cell's axon triggering reactions at its many synapses.

Eccles' intuition was that this type of neural events was the perfect recipient for the intervention of mental events in the brain. For a long period his research on the mechanisms underlying this interaction was centered on the hypothesis, inspired by Henry Margenau[315], that the mind was analogous to a quantum probability field, which has neither mass nor energy yet can cause effective action at microsites. If it could be shown that the mind acted in this way, this would allow Eccles' dualist hypothesis to overcome the accusation of violating conservation laws. In a later period, however, he proceeded to develop a more defined account together with the physicist Friedrich Beck, in which they presented a theory of how this process might take place, with precise calculations of the energy, time and distance required for quantum effects to be significant. The main claim of their theory was that the self's intentions become neurally effective by momentarily increasing the probability of exocytosis in certain cortical areas named dendrons – bundles of pyramidal-cell dentrites (the receiving ends of the neurons) that have over 100,000 synapses. The structure of these thousands of boutons "provides the *chance* for the mental intention to change by *choice* the probability of its synaptic emission"[316].

[314] Eccles, J.C. (1994), p.152.
[315] Cf. Eccles, J. C (1970) and (1989).
[316] Eccles, J.C. (1994), p.76 (emphasis added).

The dualistic stance of Eccles and Beck's theory encountered a lot of suspicion and, maybe due to that, their overall hypothesis did not have any relevant impact. However, Eccles' work must be recognized as a very serious attempt to understand how the mind might control the brain and it still remains a cornerstone for many critics of materialism. Also, his suggestion that vesicular release in exocytosis is the sort of event where the minimal quantum effects might be relevant and therefore influence the functioning of cognitive processes is still present in the work of both scientists and philosophers today.

4.6.2.2. The measurement problem and the "consciousness causes collapse" interpretation

It is usually said that the devil is in the details. It is not enough to say that consciousness can influence the outcome of a probabilistic event such as exocytosis; it is necessary to understand how that can physically happen. Even if I cannot develop an empirical theory about this, I will now try to explain briefly what is consensually known about the quantum world and how the mathematical formalism of quantum mechanics leaves room for the causal role of consciousness in the unfolding of events. My goal is not to present any uncontentious interpretation (there is no such thing in quantum physics), but rather to show how mental-to-physical interaction can fit into our scientific picture of reality without contradicting what we currently consider to be our best models of the functioning of the world.

In the world of particles at the subatomic scale, classical physics does not apply. Only once a certain measurement is made, is the system in a well determinate state; before, in general, it is considered to be in what is called a superposition of states. There is no way of knowing with absolute certainty all the information that fully characterizes a physical system: if we know precisely where a particle is located – its position in a certain spatial coordinate –, we will miss all knowledge about its momentum (in the same coordinate), and vice-versa[317]. This is radically different from what happens in the macro world where

[317] This is, of course, an extreme example of the renowned Heisenberg uncertainty relations.

we can describe accurately the complete state of an object, and where, given initial and boundary conditions, our theories can, in principle, predict the deterministic sequence of its future states (rather than the deterministic evolution of the probabilities of its being in *this* or *that* state), independently from our epistemic access to it.

Quantum mechanics was the theory developed in order to cope with the strangeness of the quantum world[318]. It is a theory that describes with extreme precision the evolution of the probability distribution of the states of a quantum system. However, it only allows us to calculate the complete set of *probabilities* of the outcomes of a certain measurement (using the famous Schrödinger equation that describes deterministically the evolution of the wave of probabilities), not their definite values. Let me make this clearer by using an example introduced by Eugene Wigner:

> Given any object, all the possible knowledge concerning that object can be given as its wave function [which] permits one to foretell with what probabilities the object will make one or another impression on us if we let it interact with us either directly or indirectly. The object may be a radiation field and its wave function will tell us with what probability we shall see a flash if we put our eyes at certain points (…).
>
> Suppose that all our interactions with the system consist in looking at a certain point in a certain direction at times t_0, t_0+1, t_0+2,…, and our possible sensations are seeing or not seeing a flash. The relevant law of nature could then be of the form: "If you see a flash at time t, you will see a flash at time t+1 with a probability ¼, no flash with a probability ¾; if you see no flash, then the next observation will give a flash with the probability ¾, no flash with a probability ¼"(…). The wave function in such a case depends only on the last observation and may be ψ_1 if a flash has been seen at the last interaction, ψ_2 if no flash was noted. In the former case, that is for ψ_1, a

[318] As is well known, the quantum world has many other "strange" aspects that are not relevant to what I am discussing here. One of them is quantum entanglement, in which two (or more) physical systems are in a correlate quantum superposition state whereby the definition of the value of a property in one of them will have non-local instantaneous causal effects over the other [or to be a little more technical: where two (or more) physical systems are in an *eigenstate* of a certain observable, but neither is in an individual *eigenstate* of that observable]. This is the phenomenon that is currently considered to be crucial to bird brain navigation, as we saw in section 3.5.

> calculation of the probabilities of flash and no flash after unit time interval gives the values ¼ and ¾; for ψ_2 these probabilities must turn out to be ¾ and ¼.
>
> (…) The important point is that the impression which one gains at an interaction may, and in general does, modify the probabilities with which one gains the various possible impressions at later stages. In other words, the impression which one gains at an interaction, called also *the result of an observation*, modifies the wave function of the system."[319]

A crucial feature of quantum mechanics if that it is a theory about measurements, not about the underlying ontology. It is a set of mathematical postulates about physics that allow us to describe probabilistically what we can predict but do not purport to explain quantum reality. Hence, the problem of measurement: what happens when the wave function of different less-than-unity probabilities collapses[320] into one single well-defined state? This question is the object of the so called different interpretations of quantum mechanics.

[319] Wigner, E.P. (1967), pp.171-2.

[320] That there is such a thing as the collapse of the wave function is already an assumption of the theory. The "collapse" can be intended in two ways: In its broader sense, it is the name given to the transition from the indeterminacy prior to measurement to the definite state that follows it. Most interpretations of quantum mechanics assume there is a collapse in this sense, which is to say they accept that there is a measurement problem. In its narrower sense, the collapse of the wave function is intended as the *instantaneous* reduction of the superposition of different states of the isolated system to one single well-defined state upon observation. In recent decades, this idea of the instantaneous reduction as a fundamental phenomenon has given way to theories that try to provide a *mechanistic* explanation of the system's transition. The most popular one is that according to which there is a process of "decoherence" in which the system's internally ordered states loose coherence (the system's component's phase angles are decoupled, which leads to the loss of its quantum properties) because of their thermodynamically irreversible interactions with a large-scale environment. However, decoherence only explains why we cannot observe quantum effects in our macroscopic everyday world; it does not allow us to predict which definite state will be selected. Also, the theory cannot give us results that are detailed and quantified, as we cannot observe nor calculate all the processes that lead to decoherence. Therefore, the fact that quantum decoherence is now the mainstream term used to refer to the transition from a quantum to a classical state does not imply that there is any consensus among quantum physicists about the matter, much less that the measurement problem is solved. In this section, I will still refer to the transition as a "collapse" because that is how the authors whose theories I am presenting intended it.

The orthodox interpretation (often called – albeit misleadingly – Copenhagen interpretation) considers that there is an *in principle* impossibility of knowing what exists in the quantum realm before a measurement is made. It thus defends a radical and unsurmountable epistemological indeterminism, grounded in our lack of epistemic access to the outcome of a measurement before it is made[321]. Thus the orthodox interpretation does not give us an account of what happens in an act of measurement, when the probabilistic wave-function collapses into a singular and well defined state. There is an instantaneous transition between the quantum (linear superposition of probability states) and classical (singular state) descriptions of the physical system, a transition to which there is no further explanation to be given by the theory itself.

There are many alternative interpretations of the quantum formalism. Bohmian mechanics, for example, states that the wave function provides only a partial description of the system, i.e. that there is at least one property in the physical world that the theory is not taking into account – namely, the actual positions of the particle – the knowledge of which would allow one to overcome the abovementioned epistemic limitations[322]. Ghirardi–Rimini–Weber interpretation explains the transition by the spontaneous reduction of the wave packet in certain particles which would initiate a causal chain of correlated reductions of probabilities to well-defined states[323]. Everett's famous multiple worlds interpretation states that for every measurement all possible outcomes become actual, each one in its own universe[324]. These are but some of the many alternative interpretations currently under debate, which have not reached a consensus and have even mushroomed in the past decades. The measurement problem is still open today.

The interpretation I believe is most interesting for the question we have been asking ourselves in this section if the one suggested by Nobel laureate Eugene Wigner, following the steps of John Von Neumann's. In the last part of his influential book *Mathematical*

[321] Note that it is not true that the orthodox interpretation of quantum mechanics has proven that there is ontological indeterminism in the quantum world. Given its nature, which makes no attempt at knowing what quantum reality is, it talks only of epistemological indeterminism.

[322] Cf. Bohm, D. (1952). This interpretation questions the idea of "collapse" in its broad sense.

[323] Ghirardi, G.C., Rimini, A., Weber, T. (1986).

[324] Everett, H. (1957).

Foundations of Quantum Mechanics (1932/1955), which contributed greatly to the axiomatization of quantum mechanics, Von Neumann enunciates explicitly the measurement problem: What is it that triggers the transition from quantum to classical modes of existence? In other words: how is it that, in a measurement process, a physical system goes from a superposition state to a "classical" well-defined state? We may want to establish it is the detection process performed by a certain measuring device, but the boundary between the part of the world that is being observed and the part that constitutes the observer is "arbitrary to a very large extent"[325]. The macroscopic apparatus that produces the measurement is a physical system, subject to quantum laws[326]. Thus, it too is a quantum system in a superposition state in need of a measuring device that may trigger the collapse. And of course any other measuring device that one might postulate could measure the first physical apparatus would also have to be included in the quantum description of the composed system [quantum object + apparatus1 + apparatus2]. Also, the body of the conscious observer that is located at the end of this chain must be included in the joint wave function that describes the composed system, for even the interaction of the photons with her retina and the following chemical changes in her brain can be included as phenomena that the subject "observes"[327]. So where does the collapse take place in this chain?

Von Neumann did not answer this question directly but he opened the door for the so called Von Neumann-Wigner interpretation that followed, by claiming that the subjects' own state is completely known to her through introspection. The state of the conscious observer before the measurement is not indeterminate.

Seven years after Von Neumann formalized this problem and claimed that the self-conscious subject was an entity which knows the state she is in before an observation is

[325] Von Neumann, J. (1955), p.420.

[326] The quantum theory's scope is not formally limited. The fact that we usually apply it only to microscopic objects has to do with the irrelevance of quantum effects in many-body systems, not with any theoretical limit concerning size.

[327] Cf. Von Neumann, J. (1955), p.419.

made, Fritz London and Edmond Bauer put forward the first explicitly subjectivist interpretation of the collapse:

> "A measurement is achieved only when the position of the pointer has been *observed*. It is precisely this increase of knowledge, acquired by observation, that gives the observer the right to choose among the different components of the mixture predicted by theory, to reject those which are not observed, and to attribute thenceforth to the object a new wave function, that of the pure case which he has found.
>
> We note the essential role played by the consciousness of the observer in this transition from the mixture to the pure case. Without his effective intervention, one would never obtain a new function."[328]

Wigner followed in 1961 and enriched his aforementioned example in order to help us understand this better: what if it were a friend of mine, rather than me, the observer of the event of a flash showing or not showing at time t?

> "One could attribute a wave function to the joint system: friend plus object, and this joint system would have a wave function also after the interaction, that is, after my friend has looked. I can then enter into interaction with this joint system by asking my friend whether he saw a flash. (…) If he says no, the wave function of the object is ψ_2, i.e., the object behaves from then on as if I had observed it and had seen no flash (…) However, if after having completed the whole experiment I ask my friend, "What did you feel about the flash before I asked you?" he will answer, "I told you already, I did [did not] see a flash," as the case may be. In other words, *the question whether he did or did not see the flash was already decided in his mind, before I asked him.*"[329]

Wigner's "friend" is a self-conscious entity whose insight is considered sufficient for her own state to be well defined. This reasoning made Wigner conclude that consciousness must have a role in quantum mechanics that an inanimate measuring device cannot have.

[328] London, F., Bauer, E. (1939), translation in Wheeler, J.A., Zurek, W.H. (1983), p.251.
[329] Wigner, E.P. (1967), p.176 (emphasis added).

It is the consciousness of the living observer that breaks the endless chain of measuring events, causing the self-triggered collapse of the system's wave function.

Even though the Von Neumann-Wigner interpretation is not one of the most popular, it is certainly compatible with the formalism and recognized as a serious alternative to the orthodox interpretation. It has also constituted the theoretical basis upon which the two above mentioned theories of the mind-brain relation were elaborated: Stapp's as well as Penrose and Hameroff's.

Neither Von Neumann nor Wigner explained *how* it is that consciousness causes the collapse of the quantum system's wave of probabilities. This is a form of psychophysical causation: how does it work? As a mathematician and a physicist, this question is clearly out of their scope. And to this day, it remains an unsolved mystery.

However, it is of the utmost importance for us to realize that they both felt that the only way to solve a theoretical impasse in physics was to reach out for the non-physical phenomenon of self-consciousness. According to them, the exclusion of the Cartesian mind from the physical realm was put to an end with the advent of quantum physics, the laws of which could not be formulated without an explicit reference to the conscious observer as a primary reality.

And as Wigner put it, in spite of our ignorance about what future science will bring:

> "[i]t will remain remarkable, in whatever way our future concepts may develop, that the very study of the external world led to the conclusion that the content of the consciousness is an ultimate reality."[330]

4.6.2.3. Why are these hypotheses relevant

Let us now try to close the circle of the related philosophical problems that have been under discussion so far: agency, free will, emergence and consciousness. First, we saw how agency requires agent-causation, and how that in turn requires a unified and

[330] *Idem*, p.169.

autonomous self. In order to see how such an entity could fit into a naturalistic picture of the world, I developed an account of emergence that defeats accusations of scientific implausibility and then argued for the irreducibility of consciousness to the material world. I then defended that the conscious self as a naturally emerging entity is the best candidate for the role of autonomous agent, and this finally led to the question how that conscious agent could effectively cause changes in her physical body.

This is where the proposals by Eccles, Von Neumann and Wigner came in. Note that there are no scientific *proofs* of psychophysical interaction (intended as a causal relation between two non-identical entities), nor evidence against it, and, given the nature of the *relata* involved, there may never be any. The arguments that are put forward in favor or against emergent dualism are based on interpretations of the ontology underlying our observations and have to appeal to metaphysical assumptions and methodological principles such as the Causal Closure of the Physical, Occham's Razor and the like. No scientist can have a final word on these matters. The Eccles and Von Neumann-Wigner proposals presented here are evidence as to how among the hot debates that have been going on in the scientific arena in the past half a century, many questions which were left unanswered can draw some light on some parallel philosophical conundrums and be amenable to some cautions speculation.

Eccles revealed to us that some neural processes are probably genuinely indeterministic and therefore vulnerable to some extra cause contributing to making one alternative more likely than another. Even if his concrete hypothesis was not on the right track, the questions that guided him are the same that guide libertarians today, and his empirical approach shows that it is reasonable to expect that neuroscience might provide us with better accounts of this process in the future. Von Neumann and Wigner's answer to the measurement problem is symptomatic of how quantum physics has an unsolved mystery at its core and how its intrinsically epistemic nature breaks the traditional separation between objective and subjective reality. Even if their suggestion of consciousness as a solution to the open problem of the transition from *probabilia* to *actualia* is not the most popular, it is still on the table, among tens of other interpretations, many of which (Everett's many-worlds interpretation, for example) are comparatively much more distant

from our common intuition. Given my argumentation showing how agent-causation is the only metaphysical approach that can vindicate our commonly held distinction between actions and unpurposive behavior, and how this requires that the conscious self sometimes steers the wheel of neural events, Von Neumann and Wigner hypothesis gains greater philosophical interest. Who knows what future science will bring, but if action is to be considered possible and free, I am betting on some developments in this direction.

Some will counter that this is no way to do science. True, but that is not what I am doing. I am doing philosophy and my arguments have led me to the conclusion that there are two mutually exclusive alternatives, none of which has been confirmed nor refuted by science:

1) There is an irreducible self who can downwardly cause our neurons to fire in such a way as to make our body move according to her will, and this implies that consciousness can affect our brain; or
2) Brain events are causally closed to any non-physical influences, which, given the argument I developed in chapter two, entails that there is no real agency in the world.

This is the point where the debate comes down to conflicting basic intuitions, and where I bet on the first alternative rather than on the latter. For all the reasons I have stated so far, it seems much more unreasonable and implausible to me to think that consciousness is identical to its physical substrate or epiphenomenal, that actions and non-actional behaviors have the same causal etiology and that libertarian free will is an illusion, than to think that the brain works indeterministically and that our conscious emergent self can affect it.

If this will imply the discovery of new entities in physics or the transition to a new way of regarding the material world altogether, so much the better. As Noam Chomsky noted[331], there is no consensual and definitive concept of what “material” or “physical” means, for if it means (as is usually intended), what is potentially describable by physical science, we

[331] Chomsky, N. (2000), chapter 4.

are definitely facing a practical problem. For example, Descartes seemed to have the physical as opposed to the conscious parts of reality as distinct as could be, and then Newton reintroduced strange occult forces into the physical realm by accepting mysterious action at a distance. More recently, quantum physics operated a true revolution on how the scientific community conceives matter and its frontiers, and its consequences are still hard to grasp today, a hundred years after its beginning. And now we are told 90% of the universe is dark matter, which by definition is everything our present natural science fails to detect and knows nothing about. How can one be sure, then, of what a physicalistic account of reality should be, and which falsities should be refuted as antiscientific ghosts in the machine?

Chomsky's argument serves to say that to take consciousness as a fundamental part of reality and to defend its irreducibility to the physical world as we currently intend it is a position forced upon us by the contingent frontiers of today's science. If in the future the "physical" will once more be expanded as to include phenomenal consciousness as well as third-person descriptions of reality, then we will not have to embrace dualism any longer.

In our present situation, however, emergent dualism *is* the best option for realists about action (whom, I contend, must endorse agent-causalism), and one that is not as demanding as usually depicted.

When assessing a theory's virtues, we must assess its explanatory power, its internal coherence and also its empirical plausibility. Remember Franklin's definition according to which plausibility is a function of how radically our current scientific knowledge would have to change for our theory's commitments to be satisfied. From what we have seen, agent-causalism does not conflict with our scientific picture of the world. Its commitments regard either metaphysical assumptions such as the Causal Closure of the Physical, reductionism or determinism, domains in which scientific knowledge is still "in its infancy"[332], such as the neuroscientific study of the self, consciousness and self-consciousness, or interpretations of open scientific problems such as the measurement

[332] Kircher, T. and David, A.S. (2003), p.8.

problem in quantum mechanics. Hence, according to Franklin's criterion, agent-causal libertarianism is not an implausible theory: things would not have to change radically from the point where they are now in order for its empirical commitments to be satisfied. True, its most popular alternatives are scientifically plausible as well; however, they would imply renouncing to our most basic conceptions of action, mental causation, personhood and free will. If we measure them up globally, we can see how agent-causal libertarianism is actually the best account available.

4.7. Consciousness and Free Will

There are open problems in both science and philosophy. Sometimes, the sets of open problems intersect, which is the case when it comes to consciousness. From what we have just seen, this concept seems to be the vanishing point towards which several unanswered questions in metaphysics (the problem of emergence), philosophy of mind (the mind-body problem), and physics (the measurement problem) converge.

In order to better understand the journey we have taken from action theory to metaphysics, to the mind-body problem, I now wish to close this chapter by pointing out that the relation between the free will debate and the problem of consciousness in philosophy is not new: it has always been very close. It is also important to note that to interpret the irreducible agent as the agent's conscious self is not to limit agency arbitrarily, leaving out all the cases in which an agent would be acting unconsciously. I do not believe there are any such cases, as I explained in the first chapter. Agency requires conscious control, independently from the requirement of a unitary self.

As O'Connor has noted, it is a "remarkable feature of most accounts of free will that they give no essential role to conscious awareness"[333]. But it has been an implicit assumption in all theories of free action and free will that for any action to be considered such, independently from the degree of deliberation involved (cf. Table 1) the agent has to be

[333] O'Connor, T. (2000), p.122

(at least indirectly) aware of what she is doing, since no one can control what one cannot even represent. This assumption is present in all the main views, from libertarians to compatibilists, as well as in the "willusionist"[334] theses that were produced and fed to the social media after Benjamin Libet's aforementioned experiments on the "illusion of conscious will"[335]. Experiments of this type have repeatedly shown that the brain can be predisposed to a certain decision a significant amount of time before the subject becomes aware of it, which led many to jump to the conclusion that people do not have free will. Even though Libet-type experiments and this way of interpreting the evidence has been harshly criticized on various fronts[336], they remain extremely influential to this day, hence the significance of their relying on the assumption that consciousness is required for free will.

Folk views also share the assumption that consciousness is crucial for freedom and responsibility, as several recent experimental philosophy studies led by Joshua Shepherd have shown. For example, according to subjects that were enquired in one experiment, agents that are behaviorally identical to human beings (for example, humanoid machines) are considered to lack free will and responsibility if "[t]hey do not actually feel pain (even when they say 'Ouch!'), they do not experience emotions, they do not see colors, and they do not consciously deliberate about what to do"[337]; in contrast, they can be considered free when they are conceived as capable of conscious experience. These considerations are quite independent from whether the context of the "zombie" is identical to the context of its conscious counterpart (that is, even if it is identically deterministic or indeterministic), or not.

This is not to say that philosophical analysis should rely on intuition. However, I believe that to contradict folk intuitions as fundamental as the ones on which our image of ourselves as humans is grounded should not be done lightly in philosophy. *Ceteris paribus*

[334] This is Eddy Nahmias' expression [Nahmias, E. (2011)].

[335] Cf. Libet, B. et al. (1983), a study which I referred to in sections 1.2 and 2.2. "The illusion of conscious will" is the title of Daniel Wegner's 2002 book, which was partially based on Libet's findings.

[336] See, for instance, Alfred Mele (2014).

[337] Shepherd, J. (2015), p.939. See also Shepherd, J. (2012).

(of course, empirical corroboration is a stronger argument than any other), a theory that confirms folk intuitions is more plausible than one which conflicts with them. This is why I think it's important to note that my account, according to which the irreducible self that is required for an action to be such can be identified with the unitary entity who is the subject of our irreducible conscious experience, meets the common intuition that free agents must be conscious.

Also in the field of moral philosophy, despite some dissident voices[338], the idea that consciousness is crucial for responsibility ascriptions is still the most common position. It is interesting to realize that it is shared even by philosophers that address consciousness as a brain function that is independent from its phenomenal quality. Neil Levy, for example, has very recently published a book on the relationship between consciousness and free will, where he argues that "consciousness of key features of our [morally significant] actions is a necessary condition of moral responsibility for them"[339]. But Levy endorses this thesis even though he does not side with its traditional defenders on what concerns attributing to consciousness unique phenomenal characteristics with which persons identify.

> "Consciousness is never more than a tiny sliver of our mental life, and the contents that happen to become conscious may not be especially significant for who we are. Consciousness is necessary for direct moral responsibility, I claim, not because of what it is, but because of what it does."[340]

According to Levy[341], the functional role of consciousness is to make information available to most of the consuming systems that compose the mind. When that information is not conscious, its contents are not "online" for many of those systems and, as a consequence, many of our beliefs and moral principles, which are distributed across many areas of the

[338] Cf. recent theories by Nomy Arpaly, Peter Carruthers and others on the possibility of responsibility without consciousness. See Neil Levy (2003) and Joshua Shepherd (2013) on the assessment and critique of these positions.

[339] Levy, N. (2014), p. vi.

[340] *Idem*, p. ix.

[341] His account was inspired by Bernard Baar's "global workspace" hypothesis [Baars, B. J. (2001)].

brain, are not accessed. For this reason, Levy claims, the actions (or responses) in such circumstances are not expressive of who we are, nor are they controlled by us.

Levy does not address the phenomenal aspect of consciousness because he believes that it is not what defines it, nor is that the aspect that is at stake when one discusses the relevancy of consciousness for moral action.

> "What is at issue in debates over moral responsibility is whether agents must have a certain kind of access to a certain kind of content in order to be morally responsible."[342]

The advantage of Levy's functional approach to consciousness as awareness is that it avoids discussions around the "hard problem of consciousness"[343], by simply moving past it and focusing on another type of phenomenon – the informational content that is available (for reasoning, for example) at a certain moment in time. Whether it has phenomenal content as well, is something that does not concern the author. This may be a useful strategy since "the thesis that [phenomenal] consciousness is ontologically irreducible to physical phenomena (and manifestly so) is a basic divide among philosophers, one that is far more intractable than the question of free will itself"[344].

For all the reasons presented in section 4.2, I cannot be agnostic about this matter. I believe it is precisely the phenomenal aspect of consciousness that which renders it irreducible and therefore a perfect candidate for the role of that emergent entity on which to ground our irreducible self. Nevertheless, Levy's thesis is useful as it shows how the functional aspects of consciousness too point towards its central position in action production.

Pace Levy, it is a fact that phenomenal consciousness exists and that it always appears together with its integrative functions. Also, the question of its adaptive value (still under debate in the field of evolutionary biology) is another open problem where the fundamental relationship between consciousness and free action gets an independent vote from. Why has nature selected this property which, as much as we can tell, seems

[342] Levy, N. (2014), p.28 note 5.
[343] Cf. Chalmers, D. (1996).
[344] O'Connor, T. (2000), p.117.

unnecessary from a strictly adaptive point of view? It is conceivable that a philosophical zombie identical to us in all its physical and mental capacities, except for phenomenal consciousness, could deal with the world just as well as we do, at a practical level. Then why was this extra first-person perspective added to our being in the world? O'Connor suggests that the connection of consciousness to free will is the key to this problem:

> "The [agent-causal] theorist can conjecture that *a* function of biological consciousness, in its specifically human (and probably certain other mammalian) manifestations, is to subserve the very agent-causal capacity [of choosing from amongst available alternatives]." [345]

In other words, maybe the biological value of consciousness lies in the fact that a zombie functions only algorithmically, since it has no personal perspective on things. That entails that it cannot choose, because choosing means to intervene in the causal sequence of brain and mental events and to author a decision oneself, and the zombie has no self. Therefore, since there cannot be agency nor freedom of choice without agent-causation, nor agent-causation without an irreducible substance with downward causal powers, consciousness was the emergent capacity selected by nature for its adaptive value of endowing the organism with that ability. By being conscious, the organism acquires a sort of unity that can then be used for acting. The conscious self is hence an emergent entity, supervenient on the body but radically distinct from it, who is simultaneously the subject who thinks and the agent who acts through that body.

We have many reasons to think consciousness and free will are intricately connected, then. My thesis is just a further argument in favor of a view that links them together.

[345] O'Connor, T. (2000), p.122. Hodgson, D. (2005) argues for a similar thesis.

5. FREE WILL AND ALTERNATIVE POSSIBILITIES

"I am morally superior to George Washington.
He could not tell a lie. I can and do not."
(Mark Twain[346])

5.1. What we have learned so far

We have come a long way. In the first chapter, I presented the foundations of my agent-causal thesis. The need for the empirical distinction between actions and non-actional behaviors to be grounded on a metaphysical distinction, together with the challenge posed by the disappearing agent objection, led me to opt for agent-causalism as the best account of agency on the market. However, as I explained in section 2.4., agent-caused actions may be more or less free. A compatibilist view about free will sees any action (intended, as I define it, as an intentional behavior caused by the agent) as a free action. Incompatibilists, instead, argue that in a deterministic world (a world in which natural laws are such that, given identical physical conditions, there can be only one possible unfolding of events), actions are not truly free. Even though a causally determined action may be voluntary insofar as it is made purposefully and according to the agent's will, the lack of alternative possibilities will make it so that the agent could not have intended it any other way. The agent can do what she wants but she cannot want anything else. And this implies that free will, as the ability for self-determination and choice, is just an illusion. In the words of Spinoza, who believed in a deterministic world, "men think themselves free, because they are conscious of their volitions and their appetite, and do not think, even in

[346] Mark Twain cit. in van Inwagen, P. (1983), pp. 63-64.

their dreams, of the causes by which they are disposed to wanting and willing, because they are ignorant of [those causes]"[347].

Chapters Two and Three have led me to embrace libertarianism (the non-skeptical version of incompatibilism) for an independent reason. In order to guarantee the naturalistic plausibility of agent-causalism, I chose emergent dualism as the best account of the mind-body relationship. I addressed the arguments usually put forward in favor of reductionism and against ontological emergentism but concluded that none of them is irrefutable and that our scientific picture of the world is perfectly compatible with the emergence of novel phenomena with downward causal powers. The reasons we have for believing that consciousness is distinct from and irreducible to matter, despite being dependent on it, provided a further argument in favor of the idea that ontological emergence exists and is present in the unified self who is the bearer of conscious properties that each one of us experiences. However, I also argued that, if one wishes to develop an account that does not question natural supervenience, one will have to posit fundamental indeterminism as a condition of possibility for downward causation. In the case of mental causation, this implies positing neural indeterminism, which I argued is a plausible hypothesis.

My belief in the existence of free agency and my contention that it is impossible without downward causation, which in turn is incompatible with determinism, led me to endorse libertarianism about free will. It is the only position that can ensure that the emergent conscious self, for whose irreducibility I provided numerous arguments, has genuine alternatives from which to choose and that, in choosing, it can be an effective cause of the actions performed. Nevertheless, I also believe that libertarianism is the best view regarding free will because of the classical reasons in its favor: because there is no true freedom in an action that the agent could not have failed to choose to do.

This chapter aims at assessing what I consider to be the main lines of the debate between compatibilists and incompatibilists: In section 5.2. I will address the most important arguments that have been put forward for and against the compatibility of free will and determinism. In section 5.3, I will focus my attention on the arguments that question the

[347] Spinoza, B. (1677), p.110.

compatibility of free will and *in*determinism, hoping to make the case for the view that not only is indeterminism not threatening for free will as it in fact can allow for an enhanced control of the agent over her action.

5.2. Free Will is incompatible with determinism

It is a commonly held intuition that free will requires the power to do otherwise, and that determinism is incompatible with it. Philosophy demands, though, that we sustain our intuitions with rational arguments and keep our unargued premises to a minimum. Let me do so, then. I will start with presenting the incompatibilist definition of free will I endorse and then show how the objections that have been put forward against it can be adequately answered:

> *Free will is the agent's ability to choose which physical or mental action she will perform, insofar as she could also have chosen to act otherwise given the exact same circumstances and laws of nature*.

This implies two fundamental conditions:

i. Alternative possibilities (AP), so that, in the given circumstances, the agent can choose between more than one possible course of action;
ii. Authorship[348] (AS), so that it is up to the agent which choice is actually made.

The second condition is the one I addressed in the first chapter of this dissertation, defending the idea that only agent-causalism can ensure it. The first condition is the one over which the compatibilist/incompatibilist divide is drawn. According to incompatibilists, in a physical world ruled by causal determinism, there is no room for alternative possibilities, which contradicts an essential feature of free action. Compatibilists can question this idea on two fronts: 1) they can argue that determinism is compatible with AP, or 2) they can object that AP is not a necessary condition for free will.

[348] This is what Robert Kane famously calls Ultimate Responsibility [Cf. Kane, R. (1998), p.33-37].

In this section, I will begin by topic (1) above: I will describe the famous "consequence argument", which has been under discussion since its first formulation (under a different name) by Carl Ginet in 1966, and still remains one of the strongest arguments in favor of the incompatibility of determinism and free will; then I will present some objections that have been put forward against it and try to show that they are ineffective. Next, I will discuss the Frankfurt-style examples that have been unremittingly used by compatibilists to show that it is not true that free will and moral responsibility require alternative possibilities of action (topic (2) above). Finally, I will present part of the recent literature on manipulation cases, which have been mostly used as a *reductio ad absurdum* of compatibilism.

5.2.1. The "consequence argument"

The most important argument in favor of incompatibilism is clearly the "consequence argument"[349], also appropriately called the "master argument"[350], which gave a formal treatment to the traditional concern (that goes back to the ancient Greeks) that determinism prevents the possibility of acting otherwise. Stated informally, the argument goes like this:

> "If determinism is true, then our acts are the consequences of the laws of nature and events in the remote past. But it is not up to us what went on before we were born; and neither is it up to us what the laws of nature are. Therefore the consequences of these things (including our present acts) are not up to us."[351]

[349] Cf. Ginet, C. (1966) and Van Inwagen, P. (1983). Van Inwagen was the one to coin the name by which the argument is best known.

[350] Cf. Kapitan, T. (2002), who refers to the consequence argument as a modern version of the "Master Argument" by Diodorus Cronos. Fara, M. (2008) also uses this term.

[351] Van Inwagen, P. (1983), p.16.

In his influential book *An essay on free will* (1983), Peter van Inwagen offered three formal versions of the argument, using technical operators created for the occasion, such as the following (used in the first and third versions of the argument, respectively):

- N_p = df. p (is the case) and no one has, or ever had, any choice about whether p (is the case)
- P_{at} (-Q) = df. agent a at t can (has the power or ability to) render Q false

Since the different versions of the argument are usually considered to be minor variations of one main line of reasoning, I will treat them collectively. As one can immediately see, both above operators rely heavily on the concept of *ability* (either as the power to render some proposition false or, negatively, as the powerlessness to determine that something is the case). Likewise, in the context of van Inwagen's informal definition presented above, to say that the past is "not up to us" means that there is nothing we can now do to change it, just like there is nothing we can do to change the laws of nature. *Can, to have the ability, to have a choice, for it to be up to one* - all these phrases will then have to be analyzed by both, defenders and critics of the argument, in order to assess the strength of its premises and its entailments.

Let us first identify the different parts of the argument. Van Inwagen's formalizations usually have seven steps, but I think we can condense the argument into a handful: three premises, an inference rule and the conclusion. All versions of the argument assume what van Inwagen regards as undeniable facts:

1) no one has the power now, or ever, to change the past
2) no one has the power now, or ever, to change the laws of nature

The third premise is entailed by the definition of determinism and the evidence that what is necessarily so cannot be changed (rule "Alpha"):

3) no one has the power now, or ever, to change the fact that, if determinism is true, it is necessary that, given the past and laws of nature, our present actions occur.

In order to proceed to the conclusion, van Inwagen puts forward a Transfer Principle (which, depending on the version of the argument at stake, can be either a transfer of power or a transfer of powerlessness principle), according to which:

> "If there is nothing anyone can do to change X, and nothing anyone can do to change the fact that Y is a necessary consequence of X, then there is nothing anyone can do to change Y either."[352]

With the help of this principle, known as rule "Beta", or of some equivalent, he can then conclude that:

4) (if determinism is true) no one has the power now, or ever, to change the fact that our present actions occur.

Since this reasoning can be applied to any agent in any situation, 4) means that, if determinism is true, no one can ever act otherwise. Given the high plausibility of the premises and the inference rules, it looks like an uncontroversial conclusion.

Let us now consider what compatibilists have to say about this argument. Despite its apparent logical solidity, the critics say, the consequence argument can be questioned both in terms of its logical validity and the truth of its premises. The main line of attack regards the possible use of a conditional analysis of abilities which might call into question the truth of rule "Beta" and also make it so that, under certain formulations, the structure of the argument will lead from true premises to a false conclusion.

What does it mean to say that, right now, I *can* clap my hands? By that sentence I can often mean that I *would* clap my hands *if* I so intended or chose or wanted or tried to. The fact is that I did not clap my hands. I am writing on my computer, and I chose not to interrupt my writing in order to prove myself a point. However, I have the clear phenomenological feel that I could have done it, had I tried. I know how to clap my hands, no one is preventing me from doing so, I am not paralyzed, and so on. All the conditions

[352] Kane, R. (2005), p.24. In response to an objection put forward by McKay and Johnson (1996), van Inwagen made a partial technical reformulation of rule "Beta" in his (2000). I will ignore that reformulation as it left the consequence argument intact and is irrelevant for the present discussion.

are in place, so I can clap my hands whenever I decide to. I have alternative possibilities of action.

But this shows only that there are (at least) two senses of the word "can": 1) a context-free sense in which it means "to have a certain skill" or "for it to be physically possible for me to do something in normal circumstances" (if I am not prevented from doing it by another agent, say, or by a disease), and 2) a context-specific sense that has to do with the concrete exact moment in which I am supposed to exercise the power (or not) to so act, given my past and the laws of nature and, of course, all the details of my present situation (up to the very last neuron in my head). Could I have clapped my hands in both these senses, or only in the first one? Both compatibilists and incompatibilists agree that the answer to this question will depend on the truth of determinism. But what they disagree about[353] is in which of these senses the abilities mentioned in the consequence argument should be interpreted.

Compatibilists say that to assume a sense of can and power that is by definition incompatible with determinism is to beg the question. And if one were to use a conditional analysis of abilities, compatibilists argue, the argument would become invalid, i.e., the first and second premises would still be true (no one can change the past or the laws of nature, no matter how hard they try), but the conclusion would be false. It would not be true that "no one has the power now, or ever, to change the fact that our present actions occur" (in the sense that one could change that fact *if* one should so try).

David Lewis (1981) and John Martin Fischer (1983) have independently argued also that there are different meanings of "could have rendered false". In a weak sense, this means only that, if one acted otherwise, then the past or the laws of nature *would have been* different in some way (though not because the agent caused them to be different – which would be the strong sense of the expression). This non-causal meaning of "rendering false" shows it to be compatible with determinism and thus, the authors say, it is sufficient

[353] Obviously, not all compatibilists think that the consequence argument is objectionable, and very few now call it into question via a conditional analysis of abilities. In this section, though, I am using the general term "compatibilists" to refer to the ones that do, such as Gary Watson, John Martin Fischer, David Lewis, Donald Davidson, among others.

to challenge the intuitive plausibility of the first two premises of the consequence argument.

Van inwagen and many other incompatibilists reacted almost with disdain to these interpretations and responded that if conditional analyses imply the invalidity of the consequence argument or the falsity of some of its premises, "so much the worse"[354] for conditional analyses, which are much less intuitively plausible than what they were trying to disprove. The only motivation for the "ad hoc"[355] move of endorsing "would… if" interpretations of "can" in the context of the consequence argument and, even more so, weak analyses of "rendering false", is a prior commitment to compatibilism.

But this is not an argument. Can incompatibilists actually demonstrate that the non-hypothetical interpretation of ability is better?

As a matter of fact, it is commonly considered that classical conditional analyses of abilities fail in general in their attempt to provide an adequate interpretation of "can" and power, independently of their application in the context of this dispute[356]. On the one hand, the counter-examples are immense. There are many situations in which we cannot adopt a "would…if" interpretation of our ability to act differently, as we simply find it impossible to try (or choose or intend) to engage in a certain behavior, like petting a dog or moving towards the deeper side of a pool (because of a trauma, say). In those cases, a conditional analysis would still attribute to us the ability of acting otherwise, even though we would in fact be unable to do what our trauma prevented us from trying to do[357]. On the other hand, it is clearly unsatisfactory to say that an agent is able to do A if she should want (or try or intend) to do A, without asking further whether she was also *able to want (or try or intend)* to do A. But this will require a further analysis of what it means to want

[354] Kane, R. (1998), p.48, quoting van Inwagen.

[355] Cf. van Inwagen (2000), pp.9-10.

[356] Cf. van Inwagen, P. (1983); Wolf, S. (1990), Berofsky, B. (2002), Maier, J. (2014), McKenna, M., Coates, D. J. (2015).

[357] Cf. Lehrer, K. (1964, 1968).

(or try or intend), and hence imply an infinite regress, unless one drops the conditional analysis at some point[358].

Usually incompatibilists argue that a "could…if" analysis is too broad, for it assumes as *sufficient* conditions that are only *necessary* for the possession of power. However, other counter-examples can be evoked for calling into question this analysis as being too narrow. Austin's golfer[359] is the most famous case: he misses a three-foot putt, even though he "could" certainly have made it, in the usual sense of the word. He has both made and missed similar putts in similar circumstances in the past. The conditional analysis is inadequate, then, for sometimes people can do something, even though they might fail to if they try.

In order to address this and other criticisms, many variations have been given to the conditional analysis in the years, none of which has gained large consensus[360]. An alternative compatibilist line of research that has been pursued recently by authors like Kadri Vihvelin (2004) and Michael Fara (2008) suggests that conditional analyses of abilities can be improved, via their very close relation to dispositions, which are grounded in the intrinsic properties of the agent. According to them, this "new dispositionalism"[361] would allow compatibilism to retain the idea that free will and responsibility require alternative possibilities but it would maintain that determinism does not prevent the agent from having them. Fara, for example, put forward the following dispositional analysis:

> "An agent has an ability to *A* in circumstances *C* if and only if she has the disposition to *A* when, in circumstances *C*, she tries to *A*."[362]

[358] Cf. Broad, C.D. (1952).

[359] Austin, J.L. (1966), p.219, n.1.

[360] For a thorough review of the literature see Berofsky, B. (2002).

[361] The term was coined by Randolph Clarke in his 2008 article "Dispositions, abilities to act and free will: The new dispositionalism". For another critical review of Vihvelin and Fara's positions, see also: Franklin, C. (2011b).

[362] Fara, M. (2008), p.848.

Fara considers that to have a disposition to A is sufficient for having an ability to A, and that to have a disposition is to have some intrinsic property in virtue of which that disposition is manifested. With this definition, Fara takes into account the possibility that an ability might be masked by external factors that momentarily and involuntarily take a certain disposition away from the agent. According to him, this view can both solve the problems faced by the conditional analysis (by avoiding the counter-examples which showed the "could…if" model to be both unnecessary and insufficient for the possession of an ability) and undermine rule "Beta".

Recall that rule "Beta" is the transfer principle which allows the incompatibilist to move from the first three premises to the conclusion of the Consequence Argument. It states that "if there is nothing anyone can do to change X, and nothing anyone can do to change the fact that Y is a necessary consequence of X, then there is nothing anyone can do to change Y either". In a case of masked abilities, an agent fails to exercise the power she has and continues to have, in virtue of some uncontrollable fact (a gust of wind in the case of the unfortunate golfer who misses the short putt). She has the ability and the opportunity to exercise it, she tries to exercise it, but she fails. According to Fara, this makes it so that while the antecedents of rule "Beta" are the case, the consequent does not follow, in the sense that the golfer retains her ability to sink the putt despite the uncontrollable gust of wind (X) and the fact that, given the wind, she could not help but miss it (Y).

I do not agree with Fara's argument, though. When we apply his dispositional analysis to rule "Beta" we get the following statement: given that there was nothing anyone could do to prevent the gust of wind, nor the fact that it would cause the golfer to miss the easy putt, then there is nothing anyone can do to stop the golfer from missing that putt. And in fact, this seems to contradict what Fara is claiming to be obvious: that the golfer *can* sink the easy putt. Anyone would agree that when we talk about the ability a certain expert golfer possesses, we believe that ability to be retained despite gusts of wind and other temporary masks that prevent the agent from exercising it on a given occasion. However, that is not what we are discussing when we assess the consequence argument and the AP condition. What concerns us in the context of the free will debate are concrete

actions in concrete circumstances and the control we can say an agent has over them under those circumstances – not the control she generally has for actions of that *type*. What incompatibilists care about is the token ability the golfer has – or lacks – to sink that short putt in that specific situation. If compatibilists respond to that worry by means of general abilities, they are just changing the subject or simply missing the point.

As Christopher Franklin noted in a critical article about the new dispositionalism, masks are commonly used as excuses. If I promised my son that I would read him a bedtime story but an eye infection is blurring my vision for a day making it impossible for me to read, I should not be blamed for failing to keep my promise. Anyone (including my son, if he is old enough to reason) would tell me that I am excused. This means that I am considered to be *unable* to read, and therefore not free nor responsible for not doing it, even though I retain my general ability as a reader. According to Franklin, this shows how "more than ability is required for freedom and responsibility"[363]. It is one thing to possess an ability or a disposition, anchored in the agent's intrinsic properties, and it is another to have the possibility of exercising it at a given time. An agent's power to act otherwise cannot be sufficient for her to be considered accountable, if it is exhausted by her intrinsic properties. She must also control whether and when she will exercise them in the particular situation at stake[364]. Don Locke could not have expressed this more clearly:

> "It is obvious that what is at issue in the free will-determinism controversy is not whether things possess powers and agents possess abilities which they do not exercise, but whether things and agents are able to exercise those powers, *even at times when it*

[363] Franklin, C. (2011b), p. 12.

[364] Various authors have defended this idea, despite the differences in their accounts and in the terminology used: Van Inwagen (1983) stresses the importance of possessing not only an ability but also the "power" to exercise it at a given occasion; Robert Kane (1998) argues for the interdependence of AP and UR (the condition of "Ultimate Responsibility", that corresponds roughly to what I call Authorship); Alfred Mele (2003) endorses the distinction between "general ability" and "specific ability"; Randolph Clarke (2008) insists on the need for "control" over the exercise of an ability; and Christopher Franklin (2011b, forthcoming b) considers that "opportunity" is as important a condition for free will as the will itself (which is based on the agent's abilities).

> *happens that they are not exercising them*. The 'can' of power and ability, in short, is not the 'can' of the free will controversy."[365]

This is why a conditional or dispositional analysis of abilities is not applicable to the consequence argument. The question the argument raises is not whether the agent has the *general* ability to act in a certain way (e.g. to make a short putt in similar situations), but whether he has the *specific* ability to do it (e.g. to sink the putt here and now). Since Vihvelin and Fara try to ground moral responsibility in just the general abilities an agent possesses, their analyses are inapplicable to the consequence argument, just like the conditional analyses of abilities were considered to be, long before their new proposal was presented.

Many other objections have been put forward against the different formulations of the consequence argument in the years, and just as many responses have been given[366]. The character of the discussion has become increasingly technical, often grounded in modal aspects that would be totally out of context in the present work.

Before we move on to the next section, it is important to remember that we have arrived at this chapter with very strong reasons for endorsing incompatibilism. Agent-causal libertarianism proved to be the best account of free action available, and the objections that are usually presented against it (which have mainly to do with its empirical plausibility) were shown to be adequately soluble. But agent-causation has metaphysical implications: substance-causation and indeterminism. This last requirement was what first took us to embrace incompatibilism. What we have seen in this section is a further argument in favor of the idea that indeterminism is required for free will, insofar as its negation would entail the negation of control.

Compatibilism is in a defensive position when it comes to the consequence argument. It has to find sophisticated ways to question its immediate plausibility. I believe that the strategies it has reached for are ultimately unsuccessful and that agent-causal

[365] Don Locke, "Natural Powers and Human Abilities" [cit. in Clarke, R.(2008), emphasis added].
[366] For a good review of the literature see Kapitan, T. (2002).

libertarianism, which I had embraced for independent reasons, is ever more cogent after this discussion.

5.2.2. Questioning the Principle of Alternative Possibilities

In the beginning of the present chapter, I presented AP as a pre-condition for libertarian free will. This means that an agent can be considered to act freely only when she could have done otherwise given the exact same circumstances and laws of nature. In the last section we have seen that the idea that the ability to act otherwise is incompatible with determinism has been challenged by many. Now we will see how the principle itself that relates AP with free will can be put into question, not so much directly, as via the questioning of the intuitive notion that moral responsibility requires alternative possibilities of action.

Authors who have taken this move usually take moral responsibility as the empirical notion of one being fit for reactive attitudes such as praise, blame, resentment and so forth[367]. Derk Pereboom uses the phrase "basic desert" and defines it as follows:

> "For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations. This characterization leaves room for an agent's being morally responsible for an action even if she does not deserve to be blamed or praised for it — if, for example, the action is morally indifferent."[368]

[367] On the notion of "reactive attitudes", see P.F. Strawson's influential paper "Freedom and Resentment" (1962).

[368] Pereboom, D. (2014a).

In his 1969 paper, "Alternate possibilities and moral responsibility", Harry Frankfurt famously argued that the idea that *a person is morally responsible for what she has done only if she could have done otherwise* (what he calls the Principle of Alternate Possibilities[369]) is false. The reason why we intuitively take it to be true is that, when we commonly (and rightfully) judge cases of coercion as cases in which an agent lacks moral responsibility, we mistakenly believe those to be just special cases of being unable to do otherwise. We think the agent is not responsible for what he has done because, being coerced to do it, he was prevented from doing something else instead. However, Frankfurt argues, this is not the right way to interpret such situations. For neither does being unable to do otherwise entail the agent's lack of moral responsibility, nor does being coerced exclude being morally responsible.

> "Coercion affects the judgment of a person's moral responsibility *only* when the person acts as he does *because* he is coerced to do so – i.e., when the fact that he is coerced is what accounts for his action."[370]

Frankfurt then proceeds to present his well-known Jones and Black cases, which are designed to show that the circumstances that prevent the agent from acting otherwise (coercion, for example) can be diverse from the ones that make him act (his own motivations, say). This should make us realize that the Principle of Alternative Possibilities that we tend to take for granted should actually be replaced by a revised formula that can primarily take into account the *reasons* that explain *why* the action was performed, rather than the fact that there were (or were not) alternative courses of action available. Frankfurt suggests the following restatement:

> "A person is not morally responsible for what he has done if he did it *only because* he could not have done otherwise."[371]

---

[369] This principle is commonly known in the literature by the acronym PAP, even though the terminology varies: sometimes it is called "Principle of Alternate Possibilities", and others "Principle of Alternative Possibilities". I favor the latter form. Unsurprisingly, Daniel Dennett chooses an entirely different name for it: CDO ("could have done otherwise").

[370] Frankfurt, H. (1969), p.833 (emphasis added).

[371] *Idem*, p.839 (emphasis added).

If, instead, the person acted as she did for her own reasons *despite* not being able to do otherwise, then she is responsible for her action. This move is very significant. Insofar as free will is the control condition for moral responsibility, the latter entails the former. Therefore, if moral responsibility can exist without AP, free will can too. This implies that if the author is right, AP can be challenged as a precondition for free will, which would vindicate compatibilism.

However, Frankfurt's cases, as decisive as they were for showing the importance of Authorship, have been questioned and discussed for more than forty years and are definitely far from constituting uncontentious counter-examples to the Principle of Alternative Possibilities. Let us look at the original[372] Frankfurt case:

> "Suppose someone – Black, let us say – wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something *other* than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way.
>
> (…) Now suppose that Black never has to show his hand because Jones, for reasons of his own, decides to perform and does perform the very action Black wants him to perform. In that case, it seems clear, Jones will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he do it."[373]

What this ingenious example helps us see is how, at least in this case, what matters most for our attributions of moral responsibility is the causal role that the agent or,

[372] Frankfurt acknowledges that Robert Nozick had independently formulated a similar example to his in lectures given years before he himself imagined them.

[373] *Idem*, pp.835-6. The original example was about $Jones_4$, since other three cases involving Jones were presented before. As a matter of simplicity, I refer to $Jones_4$ as Jones, here.

alternatively, the counterfactual intervener plays in the story, and not the fact that there are no alternative futures, given Black's presence.

In order to avoid the objection that the possibility of Black's predicting Jones' choice would beg the question of determinism, Frankfurt needed to add a footnote (no.3), in which he denies this logical misstep and introduces a detail that will be central to most objections (and counter-objections) in the Frankfurt literature in the decades to come: the possibility that a prior sign (the twitching of Jones' face, for instance) would allow Black – based on the previous deliberation process in which Jones systematically presented such a sign whenever he considered a certain hypothesis – to know what Jones is about to decide. According to Frankfurt, this sign could be indeterministically caused, but still be considered to constitute the earlier phase of a process that would unfold deterministically, once started. Nevertheless, this earlier phase would not itself be part of the decision nor of the action.

Despite this footnote (which interestingly is seldom mentioned by his critics), Frankfurt cannot avoid several objections that I will assess in the following subsections.

5.2.2.1. Prior signs and flickers of freedom

As we have seen, if Black is to know when to intervene, there must be some prior sign that can serve as a truthful marker for what Jones is about to do next. Frankfurt had already acknowledged this and so did most of his defenders and critics. The sign does not have to be external (like blushing or furrowing his eyebrows), it can be just a certain neurological pattern that a device Black placed in Jones' brain[374] can detect at t_1, indicating the decision to be made at t_2 to act in a given way at t_3.

Of course the presence of the prior sign can lead one to counter that Jones cannot be said to lack alternative possibilities altogether since, even though he is prevented by Black from acting differently at t_3, he is still assumed to be able to decide otherwise and thus to exhibit an alternative prior sign at t_1. This would be similar to Locke's example of a man

[374] Cf. Fischer, J.M. (1999) in Kane, R., ed., (2002), p.97.

who has been locked up in a room, but stays there of his own free will, since he does not know the door is bolted and he is enjoying the company[375]. The man could still *try* to leave the room, and so he does have some sort of alternative possibility. Likewise, Jones could either show a prior sign that is favorable to what Black wishes him to do, or an opposite one. This means that he does have different options at t_1.

Also, Frankfurt says the prior sign should be taken as an "earlier phase" of the action that is not itself part of it[376]. However, one can wonder whether a criterion by which such a "phase" that is not a "part" is to be identified can be given. This can be a problem since a prior sign that was somehow already part of the decision would not allow Black to intervene in time, and Jones would retain his alternatives possibilities (Black's late blockage of the action would already count as coercion).

In any case, if we take Frankfurt's suggestion for granted and assume the prior sign to be an independent behavior, then another assessment is in order: the question whether it is voluntary or involuntary. If it is a voluntary act, then it should be considered to be another action in need of Black's counterfactual control, since, by definition, Jones chooses to make it, which means he could also choose otherwise. In this case, then, the intervener should somehow move backwards in time in order to ensure Jones' appropriate sign, which would lead to an infinite regress.

If instead the sign is an involuntary one, libertarians seem to have a problem, for it will be just a mere "flicker of freedom"[377], like a random swerve, an option the agent can hardly be considered to be able to make. If the only alternative option Jones has is involuntary, then it is not robust enough for the libertarian to ground his moral responsibility on. John Martin Fischer has been defending this idea for many years, and I believe his most interesting argument is that the "flicker theorist" is mixing up "possibility" and "ability".

> "Even if another event (or set of events) occurs in the alternative sequence of a Frankfurt-type case, it does not follow that the agent has the ability (in the relevant sense) to bring

[375] Locke, J. (1689), p.238.

[376] Frankfurt (1969), note 3.

[377] This is John Martin Fischer's famous phrase.

> about this alternative event (or set of events). (…) The mere possibility of a different event's occurring does not entail that the agent has the ability to do otherwise."[378]

Several responses can be given to Fischer's argument, but, first of all, one must understand exactly what he is trying to prove. Fischer believes Frankfurt-type cases do show that moral responsibility can be attributed to an agent in spite of the lack of alternative possibilities, since she retains "guidance control" (even though she lacks "regulative control", thus lacking libertarian free will[379]). So he does not think Jones needs robust alternatives in order to be considered responsible. His point is that *if* libertarians think so, then they should also require that the alternative possibilities be voluntary. If not, they are not real alternatives Jones could choose from.

I believe a possible answer to this contention is that Jones has actually a much more robust alternative than Fischer is thinking. His alternative is *not acting at all*.

As we have seen in the first chapter, actions must be agent-caused by definition. The fact that Jones was the source of his choice rather than Black is not a causal detail that we can consider to be independent from the action he performs. We are not talking about one and the same action being brought about by two different causes in alternative situations: either Jones or Black. We are talking about one action (the one performed by Jones on his own) versus one sub-actional behavior (him performing apparently the same behavior because of Black's intervention).

In the actual sequence, Jones exhibits a certain sign at t_1 that informs Black that he has somehow set himself on a path that leads to A. In the alternative sequence, he exhibits a different sign at t_1 but the story is prevented from unfolding as it normally would because of the counterfactual intervener. Even though at t_2 and t_3 the two sequences seem to be identical, their history is different in what concerns an aspect that is essential to their very nature.

[378] Fischer, J.M. (2011), pp.260-261.

[379] See the presentation of these two types of control in section 2.4. of the present dissertation.

Margery Bedford Naylor[380] uses the example of a child who tells the truth *on her own*, instead of having to be forced to do so. Does she not deserve some praise for her action? Of course she does. Even though, given the circumstances, she might lack the alternative of not telling the truth (and she might even know that, contrary to Jones), she does choose to tell it voluntarily[381] and this fact is essential to her praiseworthiness. Therefore, Naylor argues, the child (just like Jones) is responsible for telling the truth *on her own*, despite not being responsible for telling the truth *tout court*[382]. Kane disagrees with Naylor's response to Frankfurt and asks:

> "[I]f Black does not intervene, how can Jones be responsible for doing A-on-his-own and yet *not* be responsible for doing A itself? That makes no sense. If someone is responsible for voting for the presidential candidate on his own, then he or she is responsible for voting for the presidential candidate."[383]

One must agree with Kane's objection if one is to assume an event-causal account of action. According to this view, the identity of an action is independent from its causes and the fact that, in the alternative sequence, Black is the cause of Jones' overt behavior does not render that behavior any different from Jones' autonomously caused action in the actual sequence.

On the contrary, according to an agent-causal perspective such as my own, the event caused by Black is not identical to the event caused by Jones. In order to better show this, let me use a two stage argument.

Imagine I lived in a corrupt country in which elections were tricked in such a way that my vote would always be electronically changed for a vote for a certain presidential candidate. In this scenario, I might not have the possibility not to officially vote for that candidate but, if I deliberately chose to do it, I would still be responsible for making that

[380] Naylor, M.B. (1984), pp. 250-251.

[381] *Voluntarily*, according to Naylor's terminology, is to be distinguished from *freely*, which implies the power not to do the action altogether.

[382] Like Franfurt and most authors discussing the Frankfurt-style cases, Naylor leaves the concept of Moral Responsibility unanalyzed.

[383] Kane, R. (2005), p.85.

choice of my own free will and acting accordingly. If we consider the overall history of the two events (voting for the candidate because I chose to, or voting for him because a machine changed my vote), we can easily see that one is an action that regards only me as an agent, whereas the other is a more complex process that involved the electronic device that ignored my voting intention and changed the course of events. The two events can hardly be considered identical.

Now, let us imagine a scenario in which the corruption process was somehow transferred to my head (like in a Frankfurt case), so that a device would alter my intention if I were about to decide to vote for a different candidate. If I decided to vote for the corrupt candidate without the machine having to intervene, I would be responsible for voting for him on my own, since I did it voluntarily, instead of having to have my voting intention shifted. And we can now see that, if we consider the overall history of the decision process (the connection between the overt behavior, the agent's motivations and, most importantly, the agential intervention) to be definitory of the action, then we are before two very robust alternatives: either acting or failing to act. Like in Jones' case, the "garden of forking paths" that Kane usually imagines to be lying in front of the libertarian agent[384], is actually lying behind me at the moment of my decision to vote. I can vote (behavior) *because* of the device *or* I can vote (action) on my own. These are two alternative pathways that lead both to what *only apparently* seems to be one only possible future. The unicity of this future is illusory since only in one of these cases am I truly performing an action. Therefore, Frankfurt's case for the irrelevance of AP for responsibility ascriptions has not been made.

Let us now see how this view is different from that of Fischer and Naylor's. Fischer's argument against the flicker of freedom theorists (among whom he places Naylor) is that the alternative possibilities they are attributing to Jones are not robust enough to account for his responsibility. I claim they are, since Jones' alternative of not acting at all is as robust as can be. Instead, Naylor's point is that, strictly speaking, Jones is *not* responsible for his action, contrary to what both Frankfurt and Fischer take for granted. Since PAP is

[384] *Idem*, p. 7.

not to be questioned on her view, the fact that Jones lacks alternative possibilities on what concerns the action itself takes away his responsibility for it. He is only responsible for what he has a choice about, namely, acting-on-his-own. However, contrary to Naylor, I consider that Jones is actually responsible for his action, *via* the responsibility he has for doing it on his own. If voting on my own and voting because of the activation of a Frankfurt device are to be considered fundamentally two different events (a proper action and a mere non-actional behavior), then the event that actually takes place in the world is one for which I am either responsible or not. Likewise, Jones is either responsible for the action he performs (when he does A on his own) or he is not responsible for his behavior (when he does A because of Black).[385]

But what about Fischer's accusation that flicker theorists are mixing up "possibility" and "ability"? If Jones' alternative is not acting altogether, then how can we consider him to be *able* to do otherwise? We can *a posteriori* say that he acted on his own in the sense that he was motivated by his reasons rather than moved by Black's discrete intervention, but that does not imply that he is a libertarian agent - it does not imply that, at t_1, he could have regulative control over what he was doing.

Despite its ingenuity, Fischer's objection is missing a very important qualification of the libertarian requirement for alternative possibilities. This requirement is not temporally limited to the moment of the decision or the action. If, in the past, the agent had the possibility (and the ability) to freely choose the path that led to his presently exhibiting the prior sign he did, then we can consider him to be responsible in the libertarian sense. In his libertarian theory, Robert Kane repeatedly stresses the fact that the condition of Alternative Possibilities cannot be separated from what he calls the condition of Ultimate Responsibility[386], which has to do with how we acquire the motivations we have and

[385] Other authors have pursued strategies that, in a way similar to mine, stress that the Frankfurt-agent has the power to be the *author* of his action or not: Michael McKenna (1997), Keith Wyma (1997) and Michael Otsuka (1998), to all of whom Fischer replies in (2011).

[386] "(UR) An agent is *ultimately responsible* for some (event or state) E's occurring only if (R) the agent is personally responsible for E's occurring in a sense which entails that something the agent voluntarily (or willingly) did or omitted either was, or causally contributed to, E's occurrence and made a difference to whether or not E occurred; and (U) for every X and Y (where X and Y represent occurrences of events and/or states) if the agent is personally responsible for X and if Y

become the kind of persons we are. When Dennett presents his well-known Luther example (in which he argues that Martin Luther was obviously claiming to be responsible when he proclaimed *Hier stehe ich, ich kann nicht anders* - "Here I stand, I can do no other"), in order to show that responsibility does not require the ability to do otherwise, he is ignoring what Aristotle had already made clear: that if a man is responsible for his character, then he is also responsible for the actions that outflow from it[387]. That is why Luther can be considered a libertarian agent even though, at the moment of his famous statement, he was probably really unable to act differently. The inner struggle during which he self-formed his character in several difficult decisions was what led him to that circumstance and anchored his present freedom.

Likewise, Jones' ultimate responsibility depends on the past choices which, having been made under normal circumstances – as opposed to the highly unlikely Frankfurt-style case – made it so that he would present the prior sign A rather than the prior sign B, thus determining Black's intervention, or lack thereof.

Laura Ekstrom also chooses this way of responding to Fischer's attack on flicker strategies (which she nevertheless criticizes for different reasons):

> "[T]he strongest incompatibilist view concerning responsibility and determinism does not rest its case upon the 'robustness' of alternative actions available to the agent at the time of acting, but rather upon the requirement for moral responsibility of an indeterministically generated self."[388]

This means that, even if the last moment in which Jones is given the opportunity to act (or merely behave) differently just provides him with a flicker of freedom – an involuntary neuronal pattern or facial twitch over which he can hardly be said to have control –, he can nonetheless be considered to be responsible *in a libertarian sense* if, in the past,

is an *arche* (sufficient condition, cause or motive) for X, then the agent must also be personally responsible for Y." [Kane, R. (1998), p.35].

[387] Cf. Sorabji on Aristotle's views about necessity, cause and blame [Sorabji, R. (1980)].

[388] Ekstrom, L.W. (2000), p.190.

indeterminism allowed him to make choices that were relevant for his present day tendencies and preferences[389].

But is the libertarian not presupposing indeterminism as a precondition for the safeguard of Jones' alternative possibilities of action, despite Black's presence? There is a whole line of argument, which came to be known as the "dilemma defense", that is centered precisely on how the Frankfurt cases are to be interpreted under the assumptions of determinism or indeterminism. This is what I will discuss in the next subsection.

5.2.2.2. The dilemma defense

The challenge that was put forward independently by Widerker, Kane, Ginet and Wyma[390] is this: either the Frankfurt cases assume a deterministic link between the prior sign and Jones' decision or they do not. If they do, then they will not be accepted by the incompatibilist as counter-examples against PAP, since she will consider Jones as lacking moral responsibility under determinism. If they do not, then they will not work, because if the causal link is indeterministic, the passage from the prior sign at t_1 and the decision at t_2 cannot be assured and then there may still be alternative possibilities[391]. Black will not be able to know Jones' decision in advance (and to say, as Frankfurt does, that "Black is an excellent judge of such things" is not enough to make this a coherent hypothesis). When he does learn it, it will already be too late for him to intervene in such a way as to prevent Jones from having AP, whereas if he steps in too early, the choice will be

[389] Cf. section 2.2. for the distinction between spontaneous actions (routine behaviors which can nevertheless be considered actions insofar as they are based on standing intentions and are reasons-responsive) and actions-on-the-spot (in which the agent makes fast conscious decisions). As I explained then, even though in the former type of actions the agent is not consciously intervening, her behavior is nevertheless expressive of conscious decisions made in the past. We may consider that Jones' case, in spite of Black's early intervention which prevents the natural development of an autonomous decision, is similar to a spontaneous action.

[390] David Widerker (1995), Robert Kane (1998), Carl Ginet (1996) and Keith Wyma (1997).

[391] This objection is partially what Frankfurt was aiming at in footnote 3, but he does not make it clear how the presence of a guaranteed marker for Jones' decision can avoid the charge of determinism.

determined by him, taking away Jones' responsibility. Basically, in an indeterministic world, there is no reliable prior sign. If you are a compatibilist, these examples might convince you, but if you are not, they just do not make sense.

The deterministic horn of the dilemma is usually considered to be a dead end for the Frankfurtian. Fischer, however, tried to respond to it by questioning the idea that Jones' possible lack of responsibility depends on his lack of AP. According to Fischer, since all the causes of Jones' behavior remain the same if we subtract Black, and "Jones' moral responsibility would seem to be supervenient on what has an influence or impact on him in some way"[392], then his accountability (or lack thereof) cannot depend on the fact that he cannot act otherwise. Thus, the presence of alternative possibilities appears to be irrelevant to ascriptions of moral responsibility. Once this becomes clear, the question that needs to be asked is whether determinism itself threatens moral responsibility apart from ruling out AP – a question which he answers negatively.

I do not follow Fischer in this two-step argument, though, because I do not agree that the way in which Black takes away Jones' alternative possibilities is analogous to the way in which determinism does. Black's presence removes the alternative sequence from the metaphysical space of logical possibilities without changing the actual sequence, whereas determinism's effect is much more radical. It makes it so that laws of nature predetermine what Jones' motivations will yield. They somehow "push him"[393] into the only possible future in such a way that, as we have seen in chapter three, even his ability to agent-cause his actions becomes impossible. If the unfolding of the world's history is entirely dependent on the past and the laws of nature, then there is nothing left for the agent to choose, nothing he can ultimately be the author of. Of course, these considerations can be questioned by compatibilists, but I believe the controversy they disclose is sufficient for us to be suspicious of Fischer's premises and to refuse to accept them as uncontentious in the context of this debate.

[392] Fischer, J.M. (1999), p.100.

[393] Laura Ekstrom (2000) often uses this expression when speaking about determinism.

The indeterministic horn of the dilemma, which is considered to be the most promising one for vindicating the examples' efficacy, has been given several replies, most of them as reformulations of the cases in such a way as to conceive an indeterministic causal sequence without AP and, sometimes, as to remove the prior sign altogether. I will briefly consider three approaches that I do not believe are capable of achieving this purpose, and finally I will assess Pereboom's case, which I also do not find convincing but that I think represents a deeper challenge for Frankfurt critics.

Alfred Mele and David Robb[394] have suggested the following strategy: there are two parallel sequences that operate simultaneously in Bob's brain, an indeterministic deliberative one that can lead to the desired outcome or not, and an unconscious deterministic one (that was set up by Black, the alien intervener) which functions as a backup plan. The deterministic sequence will bring about the action that Black wants Bob to perform, and preempt the causal efficacy of the indeterministic sequence, only in case Bob does not indeterministically decide on his own to do it.

This is an ingenious example but I believe it is actually contradictory. If the deterministic sequence is in place, how can we still claim that the link between the previous sign and the action is indeterministic? Bob's decision will be deterministically blocked were it to be different from what has been established, so I think the nature of Bob's deliberative process has actually been altered!

Mele and Robb have intentionally left a non-robust open alternative to Bob: he can still involuntarily become distracted and stop deciding, in which case the deterministic sequence will not be activated. Some have said that this would just be a "flicker or freedom" to which both the objections and replies presented in the last section would apply. But I actually think this is a very different situation, since the alternative would lead to a distinct action altogether (not deciding at all). However, the problem raised above still applies: although the agent can decide to A or not decide at all, he cannot decide differently. Is he still responsible for his decision? No, he is responsible for having made *a*

[394] Mele, A., Robb, D. (1998).

decision, and not for the concrete decision he made. The lack of AP has not been proven to be dispensable for the attribution of responsibility.

David Hunt[395] suggested another "blockage case" without a prior sign. In his example, he transforms the counterfactual agent into an actual one who directly intervenes in the agent's brain processes, by physically "blocking" all neural paths besides the one selected. By hypothesis, the remaining brain process is still indeterministic, although no other pathway can be activated. One might wonder, however, and similarly to the question raised above about Mele and Robb's case, how can the only remaining possibility be considered to be nondeterministic, if there is such a predetermination[396] of which neural path will be activated in the agent's brain. Pereboom asks:

> "*Could* neural events bump against, so to speak, the blockage? If so, there still may be alternative possibilities for the agent. But if not, it might seem (…) that the neural events are causally determined partly in virtue of the blockage."[397]

I think such a *determined indeterminism* is just a plainly incoherent notion, so I will turn directly to Eleonore Stump's proposal[398], which seems more promising. In order to develop her argument, she assumes an identity theory according to which a mental event is identical to the completion of a series of neural firings. If the neural sequence is interrupted halfway, it will acquire no meaning at all at the mental level and so there will be no choice or decision. If we assume the neural sequence to be indeterministic, then we can interpret Black's intervention in Frankfurt cases as the prevention of there being any decision whatsoever. There would be no mental act in Jones' mind.

This is a very interesting argument, I believe, for it relies on the empirical fact that the mental-physical relation is a one-to-many correlation, which in fact clears out some of the doubts that these examples sometimes raise. But I do not see how this can help the compatibilist case. If neural events are caused indeterministically and Black interrupts a

[395] Hunt, D. (2000).

[396] Kane refers to Hunt's example as a case of predetermination or even predestination in (2005).

[397] Pereboom, D. (2002), p.116.

[398] Stump, E. (1996, 1999).

sequence that had already started at the physical level, then Jones had already exercised his power to choose. The fact that, at the mental level, Jones cannot consciously begin to make a choice (for the mental event we call a choice does not happen if the neural sequence on which it supervenes does not come to completion) is metaphysically irrelevant at this point. The libertarian might endorse an identity theory of the mind-brain relation and not be committed to the thesis that a free action has to be made consciously. What matters in this case, is that some events involving Jones, as a psychophysical system, lead to the neural sign that triggered Black's failing to intervene – but might have led to some different event that would have made him enter the scene instead.

For a thesis such as mine, according to which the agent's conscious self is the one who authors the free decision, the response to Stump's case is less straightforward. Consciousness has a natural supervenience relationship with the brain. However, if the conscious self is an emergent entity, it is endowed with non-derivative causal powers whereby she can downwardly influence neural events. That influence takes place in time: the conscious self at t_1 causes certain indeterministic neural events at t_2. Let us take the event at t_2 to be the one that Black prevents from unfolding. Now remember that there is no immediate conscious state at t_2, since there must be a complete sequence of neural events for any conscious mental state to emerge, which in this example we take to be the sequence of events happening from time t_2 to time t_3 at the neural level. This means that at time t_2 the agent has no consciousness of having made a decision. However, at t_1 the agent's conscious self had already been the cause of the initiation of the neural sequence at t_2. Even though no decision was made, there was a causing that Black did not prevent. And if we were to push back his intervention to time t_1, then the agent-causal libertarian may simply go further back and say that at time t_0 there was a downward causal process whereby a previous influence of the self on its neural substrate led to the neural state the agent is in at t_1. Ultimately, Stump cannot develop her reasoning without falling prey to an infinite regress.

The last indeterministic Frankfurt-type case I would like to assess is Derk Pereboom's. It concerns Joe, a libertarian agent who is considering whether or not to claim an illegal tax deduction.

> "Crucially, his psychology is such that the only way that in this situation he could fail to choose to evade taxes is for moral reasons. (…) In fact, it is causally necessary for his failing to choose to evade taxes in this situation that a moral reason occur to him *with a certain force*. A moral reason can occur to him with that force either involuntarily or as a result of his voluntary activity (…). However, a moral reason occurring to him with such force is not causally sufficient for his failing to choose to evade taxes. If a moral reason were to occur to him with that force, Joe could, with his libertarian free will, either choose to act on it or refrain from doing so (without the intervener's device in place). But to ensure that he choose to evade taxes, a neuroscientist now implants a device which, were it to sense a moral reason occurring with the specified force, would electronically stimulate his brain so that he would choose to evade taxes. In actual fact, no moral reason occurs to him with such force, and he chooses to evade taxes while the device remains idle."[399]

After presenting the case and highlighting its merits, Pereboom explains that the nonoccurrence of a reason does not ensure Joe's tax evasion, since it is always possible that a reason will *still* occur to him. The future is always open until the intervener comes into play, but he will not, not until a moral reason occurs to Joe.

Notice that this case is able to avoid the accusations of incoherence of some of the previous ones, for the indeterministic nature of both the actual sequence and the alternative one (the one in which the intervention takes place) seems to be assured. Neither the occurrence of a moral reason is sufficient for Joe to decide not to evade taxes, nor does the nonoccurrence entail the opposite decision. Because of this, Pereboom's case purports to have succeeded in creating a third alternative to Widerker et al.'s dilemma: one in which the agent's decision is inevitable given the prior sign, but not because of the deterministic nature of the situation. So, if the counterfactual intervention is not triggered by any prior sign, the agent is responsible *and* he could not have acted otherwise. The fact that the prior sign is a necessary but insufficient condition for the action is what, according to Pereboom, makes his case work.

[399] Pereboom, D. (2002), pp. 118-9.

It is interesting that Pereboom leaves the voluntary versus involuntary nature of the prior sign (the occurrence of a moral reason) unsettled. Let us flesh out these two hypotheses. If the occurrence of a moral reason is voluntary, I believe the case fails, for that could already be considered to be a robust option onto which to anchor Joe's responsibility, prior to the neuroscientist's possible intervention. It is true that, given the occurrence of the moral reason, the decision might still have been different, but responsibility for an action does not depend on the certainty of success of the voluntary steps that lead to that action. So Joe is responsible *plus* (and the libertarian could say *because*) he has AP.

On the other hand, if the occurrence of the moral reason is just a random event, Joe's responsibility will depend only on the step that follows this event. However, according to the case's description, if a moral reason does not occur to him, provided no one intervenes, Joe will decide to evade taxes. This means that his inclination to evading taxes is sufficient to ensure that, in that situation, he does. Joe cannot act otherwise, under those circumstances. Now note that the random possibility that a moral reason might still come to his mind does *not* constitute an alternative possible action, but just the physical possibility, given causal indeterminism, that the *circumstances* might be different. This is definitely not what a libertarian is after. And besides, at the moment of decision, this is not a possibility any more. And if, alternatively, a moral reason did occur to him involuntarily, the device would be triggered and his action would be coerced.

Hence, my main response to this very ingenious example is that I do not believe it to be describing a libertarian agent in the first place – even though Pereboom explicitly qualifies Joe as such. He says: "If a moral reason were to occur to him with that force, Joe could, *with his libertarian free will*, either choose to act on it or refrain from doing so" (emphasis added). But if that happened, then the neuroscientist would intervene, which would prevent Joe's libertarian free will from being exercised. What makes Joe responsible is his free will, and what makes his choice free is his power to act differently in the same circumstances. Could he retain this power in spite of the neuroscientist's presence? No, he could not, since, at the moment of decision:

a) if a moral reason has not yet occurred to him, he does not have this power in the first place

b) if a moral decision has occurred to him, then the counterfactual intervener will have compelled him to evade taxes, taking away his freedom.

I think we can now move towards a conclusion. The dilemma defense put forward by Kane, Widerker and others constitutes a powerful objection against the Frankfurt cases' coherence. Their dependence on a prior sign that can be a reliable marker for the alien intervention without simultaneously determining (or being a symptom of the previous determination of) the agent's decision makes them susceptible to criticism, for that is not an easy hypothesis to make sense of. I believe the several reformulations of Frankfurt cases that were presented as a response to this problem, albeit extremely ingenious, either fall prey to the same objections as the original one, or to new ones. Therefore, I contend that they do not manage to prove that the libertarian idea that alternative possibilities are required for free will and responsibility should be called into question.

5.2.2.3. The irrelevance of Frankfurt cases for agent-causalism

There is one more brief point I need to make before we turn to the manipulation arguments that have been actively presented as an objection against compatibilism. The point is this: the arguments revolving around Frankfurt-type cases, despite their enormous importance in the philosophical debate on free will over the last fifty years, are ultimately irrelevant for an agent-causal thesis such as mine.

I believe Frankfurt-type cases do not succeed in proving the falsity of the Principle of Alternative Possibilities and I have tried to show why that is so. However, even if they did succeed, that would not have been a problem for my account, as the main argument I have for defending libertarianism is that indeterminism is necessary for the agent's irreducible self to have downward causal effects over her bodily movements. Given the assumption of natural supervenience, if determinism were true of the world, every physical event would be fixed by the past and the laws of nature, and there would be nothing left for the emergent self to cause. Therefore, even if moral responsibility (and thus free will) did not require the metaphysical condition of alternative possibilities,

agency would still require it, for different reasons. Agent-causalism would nevertheless have to side with incompatibilism.

5.2.3. Manipulation arguments

Usually compatibilists and other critics complain that libertarianism is a very "expensive" philosophical position. It entails several empirical commitments such as indeterminism or, in my case, an irreducible self, which are open to the accusation of being implausible. However, some recent thought experiments have been designed to show that compatibilism comes at a high price as well. This means both that committed compatibilists have to be aware of what their position implies (and see if they are prepared to accept it) and that agnostics have to think better when weighing up alternative accounts regarding free will.

These thought experiments typically describe a multiple-case scenario in which one or more manipulation cases are shown to be different from an ordinary deterministic case in ways that are not relevant to responsibility ascriptions. The characters which are judged to be unfree and not morally responsible in the former cases, should therefore be judged in the same way in the latter.

Pereboom was the first to present this sort of multiple-case argument in 1995. His main character, Mr. Green (who later became Professor Plum[400]), decides to kill Ms. Peacock (later called White) for egoistic reasons, and in doing so meets all main compatibilist requirements for moral responsibility:

- his act "is caused by desires that flow from his 'durable and constant' character", though they are not irresistible (Hume);
- his first-order desire to kill Ms. Peacock conforms to his second-order desires "in the sense that he wills to murder her and wants to will to do so" (Frankfurt);

[400] Cf. Pereboom, D. (2007).

- he is “reasons-responsive” in the sense that his desires are modified by rational considerations (Fischer and Ravizza)
- he retains the capacity “to grasp, apply and regulate his behavior by moral reasons”[401] (Wallace), even though in the present situation his egoistic reasons are stronger.

By listing all these non-mutually-exclusive properties, Pereboom wants to make sure that Green possesses the agential powers that compatibilists claim to be sufficient for moral responsibility (which, on their view, would entail free will). Then he proceeds to describe the four scenarios in which his character decides to kill Ms. Peacock.

In the first case, Green is locally manipulated by neuroscientists who can produce his reasons by radio waves. His reasoning process is what brings about his decision, but that process is externally controlled by the neuroscientists.

Since many compatibilists might object that Green’s obvious lack of responsibility in this case is due to his being directly manipulated, Pereboom moves on to a second case. Now Green is a regularly functioning person who was created and programmed by neuroscientists in such a way as to weigh reasons for action exactly as he does. His reasons-responsive process that is active right now was therefore causally determined to take place in circumstances such as these.

It seems unprincipled, Pereboom says, to retain the second Mr. Green morally responsible for killing Ms. Peacock, just because the time lag between his programming and the action is longer than in the first case scenario. The difference between the two scenarios is not sufficient (as it seems irrelevant) to account for any difference in responsibility attributions, for what makes us judge the first Mr. Green as non-responsible is most probably the causal determination of his action by factors beyond his control, and this is equally present in the second case.

In the third scenario, Mr. Green becomes even more similar to an ordinary person under determinism: there are no invasive neuroscientists in the story. Instead, his character and

[401] All the foregoing quotes are from Pereboom, D. (2007), p. 94.

reasoning principles were determined at a very early age "by the rigorous training practices of his home and community" aimed at making him a "rational egoist" [402].

In case 3, just like in cases 1 and 2, Mr. Green fulfills all the pre-requisites of a compatibilist responsible agent. So whatever might make one judge this third version of him as responsible, while judging the others differently, would have to rely on something other than these conditions. What could it be? Again, Pereboom argues that causal determination by factors beyond the agent control is what explains Green's lack of responsibility in the second case and, forcefully, in the third one as well.

What happens when we move on to the ordinary scenario? In it, "physicalist determinism is true"[403] and Green is just a rationally egoistic normal human being. The main difference between this case and the other three is that now the causal determination of Green's behavior is not brought about by other agents. But if we were to reformulate the first cases, in such a way that the generation of Green's reasons would not be induced by neuroscientists or other agents, but by a spontaneously created machine, unconscious and unpurposive, our intuition about his lack of responsibility would likely persist. Pereboom's conclusion is that Green's exemption from responsibility in the first case, which is grounded in the fact that his action results from a deterministic causal process that traces back to factors beyond his control, generalizes to normal cases in a deterministic universe.

Pereboom's strategy is clever: by making it clear that these four cases are similar *in what concerns* aspects that are most relevant for responsibility attributions (which makes it so that this series of cases is not a sorites), he strongly pushes agnostics into choosing the incompatibilist view according to which causal determinism is responsibility-undermining. At the same time, he forces compatibilists into accepting that if the agent in the fourth scenario is to be considered responsible, then so should his other counterparts. This implies that compatibilism brings with it the heavy load of attributing responsibility to manipulated agents, which seems to contradict our most basic intuitions.

[402] Pereboom, D. (1995), p.24.
[403] *Idem*, p.25.

Michael McKenna, a renowned compatibilist, welcomes this move and argues that we should, in fact, foreground our intuitions about normal scenarios over those that the manipulation cases suggest:

> "[The compatibilist] should fix, not upon hidden causes, but upon the sorts of agential properties that typically serve as a basis for ascribing responsibility. Once it is established that actions issuing from a (possibly) naturally determined agent invite certain sorts of evaluations in terms of responsibility, one can then hold that actions issuing from an appropriately manipulated agent should be evaluated no differently."[404]

So McKenna's "hard-line reply"[405] to the manipulation argument is, basically, that if compatibilism implies that a manipulated agent must be judged as responsible, so be it! The conclusion to which these examples point varies, he says, according to the elements to which one draws greater attention, when assessing the four cases: agential properties versus the hidden causes of action. If we focus on the former, we will move from the attribution of moral responsibility in the fourth case to its attribution in the previous case, all the way down to the first one. Alternatively, if we focus on the latter, we will move in the opposite sense, as Pereboom intended us to. So we are before a stalemate which adds little to the debate[406].

However, as we saw at the outset, Pereboom was careful enough to flesh out the details of Green's agential properties. What the incompatibilist recommends, he says, is that we draw equal attention to them *and* to the hidden causes made salient by these examples. What these cases purport to show is precisely that the compatibilist conditions are necessary but insufficient to ensure the agent's responsibility: the absence of causal determination by factors beyond her control is also necessary.

[404] McKenna, M. (2005), cit. in Pereboom, D. (2005), p.241.

[405] McKenna calls the type of counterarguments that question the idea that a manipulated person is not morally responsible for her action "hard-line replies", as opposed to the "soft-line replies" that question the premise that manipulation cases are no different in any relevant respect from normal cases of agency under causal determinism [cf. McKenna, M. (2008)].

[406] This is also Fischer's position in (2011).

I believe McKenna is actually ignoring the disturbing consequences of accepting the validity of this argument. If compatibilism implies that an agent such as Mr. Green in the first or second scenarios is morally responsible for what he does, does this not mean that this position loses any credit as a serious and plausible view on free will? A recent experimental philosophy study conducted by Eddy Nahmias and colleagues[407] has shown that the presence of actual manipulation is crucial for judgements about the lack of free will. Subjects in this study were presented with two scenarios: both described a future reality in which technological developments have reached the point where neuroscientists can predict with 100% accuracy the full behavior, decisions and actions of an agent, Jill, who accepts to wear a brain scanning device for a month. However, while in one scenario, "the neuroscientists cannot do anything to change brain activity and hence they cannot directly influence thoughts and actions", in the other they "can even use this technology to alter a person's decision by altering the person's brain activity without the person being aware of it"[408]. The subject is the same in both scenarios, thus her "agential properties" do not change. What changes is the fact that she is the sole cause of her actions in one case, while her brain activity is altered by an external agent in the other. The results of the experiments were clear: 91,7% of the subjects did not perceive the possibility of prediction as threatening to free will in concrete decisions (e.g. voting for a presidential candidate) in the "Cannot Manipulate" scenario, while 88,1% retained that Jill lacked free will whenever her actions were manipulated in the "Can Manipulate" scenario[409]. To deny this common intuition is to underestimate the importance people give to the role of the manipulator.

Nevertheless, time and again scientific knowledge and philosophical reasoning have challenged folk intuitions, which offer no guarantee of truth and are very often contradictory. McKenna's strategy is precisely to show that the incompatibilist intuitive

[407] Cf. Nahmias, E., Shepard, J., Reuter, S. (2014).

[408] These quotes are part of the text the subjects of the experiments were presented with.

[409] In the study, there are three different experimental scenarios (each with the "Cannot Manipulate"/"Can Manipulate" variation). In all three the results confirm the hypothesis, but the values I refer to in the text are those of experiment 1, as reported in table 3 [Nahmias, E., Shepard, J., Reuter, S. (2014), p.511].

reaction to Pereboom's case 1, in which Green is directly manipulated by neuroscientists, is unfounded.

In fact, folk intuitions would count for nothing if, upon reflection, we should find the manipulator's intervention to be innocuous, as McKenna claims it is:

> "The compatibilist needs to make clear that once the manipulation is so qualified that all an agent's current time-slice compatibilist-friendly structures are properly installed through a process of manipulation, then the role of the manipulator begins to shrink into the background"[410].

McKenna rightly notes that manipulation cases are non-starters if the "compatibilist-friendly agential structure" is not in place in the first scenario (the one on which the whole argument depends). Pereboom's four-case argument, however, has the merit of assuring that it is, especially if one interprets the moment-to-moment intervention of the neuroscientists' radio waves as not being too invasive (in McKenna's words, for Pereboom's argument to work, the manipulation in the first case must be a sort of "causal prosthetic"[411]). The team of neuroscientist must intervene in such a way as not to preempt Green's agency:

> "[Green] must have an internally coherent and properly causally integrated mental life. His memories about past considerations must be able to inform and causally influence his current deliberations. And he must be causally linked to the external world in the proper way. If a bus is careening along out of control ready to hop up on the sidewalk and crush him, he is able to respond to those facts and leap from danger, and so on."[412]

I agree that a more invasive interpretation would prevent the argument from being effective, insofar as compatibilists would not accept the premise that there are no relevant differences between the cases. However, I disagree that a causal prosthetic is not responsibility undermining. Even if the agent retains her agential properties, she cannot use them autonomously whenever they are hijacked by the manipulator. Green's

[410] McKenna, M. (2005), cit. in Pereboom, D. (2005), p.241.
[411] McKenna, M. (2008), p.150.
[412] *Idem*, p.149.

escaping from a bus is an action of his, but killing Ms. Peacock is not. Hence, he is responsible for the former but not for the latter.

According to Pereboom's description of case 1, the neuroscientist's intervention takes place "before [Green] begins to reason about his situation, thereby causing his reasoning process to be rationally egoistic"[413]. His first-order desires conform to his second-order desires because his process of reasoning "by which his desires are modified and produced" is "directly manipulated"[414]. All this is made in such a way that his decision is not compulsory, it fits his character, etc. But insofar as it is caused by a prosthetic reasoning from the start, Green cannot be considered himself to have brought it about. He was just the pawn in the game.

Therefore, even if the agent's mental structures are in place, the fact that their content is not autonomously formed but rather put into her by some other agent entails that she is not in control of her present decision nor of its future consequences any more. The burden of proof is still on compatibilists to show that all four cases Pereboom described are cases of free and responsible action, rather than four cases in which free will is lacking.

A similar manipulation case was put forward by Alfred Mele[415]. Mele's goal, however, was to use his manipulation argument in order to show how determinism is irrelevant for free will. His case is the following. Diana, a supremely intelligent being, creates Fred because she wants him to do something which, given what she knows about the state and laws of the universe and the desires and values she implants him with, he will not possibly fail to do a year later. He will decide to do it on the basis of rational deliberation (he is an autonomous agent, according to Mele's own defined conditions for autonomous agency), but that process of decision production had been foreseen and intended by Diana when she created him. If the universe is deterministic, even though he could (in a compatibilist sense of 'could') change his values, desires and decisions, "there is no chance that he will"[416]. Mele's point with this example is much similar to what McKenna objected to

[413] Pereboom, D. (1995), p.23.
[414] *Ibidem*.
[415] Mele, A. (1995, 2006).
[416] Mele, A. (2006), p.185.

Pereboom: the fact that Fred was created as he was with the purpose of fulfilling Diana's wishes does not change the nature of his autonomy, which is grounded in personal characteristics of his, such as self-control, his capacity for informed deliberation and the lack of compelled motivational states.

But surprisingly for Mele, the compatibilist Tomis Kapitan countered (in a much typically incompatibilist way) that Fred's case might actually be used to falsify Mele's own conditions for free action, for Diana endowed him with pro-attitudes which will deliberately cause him to act in a certain way under certain circumstances, "in much the same way that designers of robots program the responses of their machines to various stimuli"[417]. Given this consideration, Mele asks what might justify the fact that Norm, another character who is just like Fred albeit having been born naturally, is judged by compatibilists to act freely in a deterministic world, except for historical factors which a traditional compatibilism does not usually give much relevance to[418].

In order to avoid considerations based only on the argument that, on the basis of externalist theories of mental states, a fully grown instantaneously created and mentally developed Fred is simply impossible, Mele put forward a very interesting variation on the original case in his 2006 book, *Free Will and Luck*[419]. The new "zygote argument", which has inspired many discussions lately, describes also a deterministic universe in which the powerful goddess Diana decides to create a person who will eventually do certain things that she wishes the person to do. However, in this case, she manages to mix up some atoms in the appropriate way so as to implant a zygote in a woman named Mary. Given Diana's total knowledge about the state of the universe and natural laws, the zygote will successfully become Ernie, the person she intended to create, who will act in a certain way thirty years after his creation, just as planned.

> "Thirty years later, Ernie is a mentally healthy, ideally self-controlled person who regularly exercises his powers of self-control and has no relevant compelled or coercively

[417] Kapitan, T. (2000), "Autonomy and Manipulated Freedom", cit. in Mele, A. (2006), p.187.

[418] Note that Mele believes compatibilists should opt for a quite history-sensitive approach to moral responsibility [cf. Mele, A. (2013)].

[419] Mele, A. (2006), p.188.

> produced attitudes. Furthermore, his beliefs are conducive to informed deliberation about all matters that concern him, and he is a reliable deliberator."[420]

What do our intuitions tell us regarding Ernie's responsibility when he acts? Does Kapitan's analogy with a robot apply to this case too? At this point, Mele presents us Bernie, who has exactly the same properties as Ernie. In his 2014 *Dialogue on Free Will and Science*, Mele even fleshes out the details better: Diana has written a novel with a character she now decided to make real; while in the first case scenario her total knowledge made her deduce that, in order to create Ernie, she had to mush some atoms together in a certain way as to produce his zygote *ex novo*, in the second case she realized that there was a naturally conceived baby named Bernie who was such that he would do all the things that her novel's character did, which allowed her to sit back and relax, without having to actually intervene in the natural order of the world in order to create him[421].

If one assumes a compatibilist stance, then Bernie should undoubtedly be considered a free and responsible agent. But if Bernie is just like Ernie, both in what concerns his properties and abilities and the actions he will perform in the future, then why should the historical aspects of his creation make us judge him any differently?

These considerations led Mele to present an incompatibilist argument that he believes can be more successful than Pereboom's four-case argument, even though he will not embrace its conclusion, given that he is agnostic about the truth of its first premise:

> "1. Because of the way his zygote was produced in his deterministic universe, Ernie is not a free agent and is not morally responsible for anything.
>
> 2. Concerning free action and moral responsibility of the beings into whom the zygotes develop, there is no significant difference between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist in a deterministic universe.

[420] *Ibidem*.

[421] Mele, A. (2014a), pp.15-18.

3. So determinism precludes free action and moral responsibility."[422]

Some critics like McKenna or John Martin Fischer[423] might simply deny premise one, while retaining premise two. Intuitions are crucial here: incompatibilists will tend to take the first premise to be true and compatibilists (pace Kapitan) will tend to consider it to be false. What should be assessed is whether their intuitions about this particular case can be defended without relying on their previous independent intuitions about the possibility of free agency in a deterministic world.

I will try my best to defend the zygote argument without begging the question now. First, premise 1 states that the reason why Ernie is not free and responsible for his actions has to do with his having been created by Diana. I agree. If it were not for that fact, it would not be possible to assert his freedom and responsibility independently from compatibilist or incompatibilist considerations. Like in a case of compulsion or direct manipulation, determinism becomes close to irrelevant here and what primarily prevents the agent from being free are the specific contours of his case.

Ernie's creation was not casual, it was intended as a way to fulfill Diana's plan regarding events that she wanted to make sure would happen. She had to put together his atoms into his DNA and all the parts of his cells in such a way that the resulting zygote would bring about the desired outputs. As the product of these ingredients and this "recipe", Ernie is like an actor who has been given this role to play. When he reasons, decides and acts, he is just fulfilling a previously designed plan. Kapitan is right: this sounds exactly like a computer engineering process and the result is analogous to an adequately programmed science-fiction robot.

What could a compatibilist argue the difference between Ernie and a future reasons-responsive computer to be? By hypothesis, the difference cannot be the robot's deliberative capacities, for the way in which I am using Kapitan's analogy supposes a future computer to be able to reason just like a human person. But the compatibilist might just bite the bullet: well, if the robot reasons like a human person, it will be just as free,

[422] Mele, A. (2006), p. 189.
[423] Fischer, J.M. (2011).

for what endows an agent with free will and responsibility, on our view, are precisely her intrinsic capacities to decide on the basis of reasons.

I believe the heavy price of defining free will only on the basis of intrinsic properties becomes very clear once we follow such a dialectic way of reasoning. For if a programmed robot is to be judged free and responsible, then it will probably become hard for many agnostics to accept the compatibilist premises that have led to this conclusion. The reason for my saying this is not the robot's silicon constitution and deterministic functioning – as that may not be a problem for agnostics – but rather the fact that its responses to the stimuli are directly programed by designers, which is clearly problematic when it comes to control. If the designers have set an algorithm that determines that the robot will react to the stimulus A by doing B, then how can one say that the robot is the one is control of its performance when it comes to reacting to A? There is no room for agency here, as the robot is merely following rules.

Maybe it could be easier for compatibilists to argue for the falsity of the second premise[424]. Why should one consider that the differences between Ernie and Bernie (or Fred and Norm) are not relevant to their free will and moral responsibility? Why should historical important details such as Ernie's purposeful production not be significant? For instance, one could contend that the fact that Diana is the machiavellian mind behind Ernie's life, behind everything he ever did and all the consequences of everything that happened because of his coming into the world, then *she* is the one to be ultimately blamed or praised for all these things, not him. None of it would have happened if it were not for her, and he is only an intermediate link in this causal sequence, someone who has been "trapped" into being a willing puppet in Diana's hands [425]. And if we choose to define free will only on the basis of responsibility, then Ernie is not free either. On the other hand, Bernie, whose story is just the unfolding of regular and natural events, can be

[424] Cf. Barnes, E.C. (2013) and Waller, R.R. (2014).

[425] Cf. Todd, P. (2013), p.194.

appropriately considered blameworthy or praiseworthy for what he does, according to compatibilist conditions. And if he is responsible, then he must possess free will[426].

This is an interesting suggestion but I believe it is based on an illegitimate swapping of levels of analysis. A compatibilist that choses to challenge the second premise of Mele's zygote argument will consider that there are responsibility-relevant differences between Ernie and Bernie (or Ernie versus anyone living in a deterministic universe) whereby they should be judged differently. But what are the criteria for these different judgements?

When the compatibilist considers that Ernie's lack of free will and responsibility, unbeknownst to him, is to depend on the fact that Diana is ultimately responsible for what he does, she is talking about responsibility as something that can be attributed to an agent by an observer in function of contextual circumstances, not only in function of the intrinsic features of the action production event. Also, on such an analysis, free will ceases to be the power an agent has to control his action, and becomes a derivative concept, one that corresponds to a capacity whose possession can be inferred via the external assessment of who is responsible for the action. Since identical agents in identical circumstances can have it or not, depending on how someone else might be considered to be more accountable for the action than any one of them, this capacity seems not even to be a feature of the psychological structure of the agent. I suggest that this type of analysis can be classified as *externalist.*

On the other hand, when the compatibilist considers Bernie to be responsible on the basis of his mental health, self-control and deliberative capacities, he is moving to an *internalist* level, one which has to do with Bernie's psychological process of intentional decision rather than with the broader context of that decision (this is, in fact, what allows compatibilists to ignore the threat of determinism). This does not mean that the compatibilist does not take into account certain historical factors such as Bernie's upbringing (since a possible trauma might be responsibility diminishing). However, these historical factors would have relevance only insofar as they might impair some of the

[426] This interpretation of the Ernie/Bernie difference is suggested (though not responded to) in Mele, A. (2014a), pp. 17-18.

actual capacities Bernie possesses at the moment of decision. In Mele's story, however, both Ernie and Bernie decide on the basis of "sheddable" values, which means that they are able to change their values if they put their minds to it.

So if one were to question premise 2 on the basis of the above considerations, one would be shifting from an internalist analysis by which Bernie should be considered responsible, to an externalist one by which Ernie should not. I believe this kind of move can hardly be considered legitimate. Either both Bernie and Ernie are responsible, based on internalist considerations, or none of them is, if one assumes an externalist point of view. My position is obviously the latter. I believe historical factors are crucial insofar as they could undermine the agent's ability to be the ultimate author of her action – the control condition for moral responsibility. In Ernie's case, that authorship belongs to Diana, as she was the one who set him up for the actions he is predetermined to perform. In Bernie's case, Diana did not interfere in his making, but blind natural causes ended up determining that he be just as effective a tool for her purposes as Ernie.

My conclusion, then, is that manipulation arguments such as Pereboom and Mele's help the case of the incompatibilist insofar as they make clear that compatibilist conditions are insufficient to prevent the attribution of free will and responsibility to agents one would likely let off the hook. The subtle transition between cases in Pereboom's examples, just like in Mele's zygote argument, allow us to see how the hidden causes in a deterministic scenario are very similar to distant manipulation by external agents, which means that they are much more responsibility-undermining than what the compatibilist wishes to admit.

5.3. Free Will is compatible with *in*determinism

The arguments that have been presented in the previous section, aimed at defending the incompatibility between free will and determinism, as well as at attacking rival views, are not crucial for my account. As became clear in chapter 3, given the structure of my reasoning which is based on the need for an irreducible agent with downward causal

powers, I have independent reasons for endorsing an incompatibilist account of free will. Therefore, the consequence argument, the incompatibilist responses to Frankfurt cases, and the more recent manipulation arguments, albeit interesting and enriching, are inessential to my libertarian stance.

The arguments that will be presented in the present section, however, could, if sound, shake the foundations of libertarianism and force me to embrace hard incompatibilism[427]. In fact, it is not enough to postulate neuronal indeterminism as a precondition for downward causation from the self to its brain; human free will requires also that the agent be able to decide for reasons while retaining the ability to choose otherwise, under the exact same circumstance, i.e., given the complete state of her mind (including all the reasons that are present to it) at the moment of decision. Many have complained that it is impossible to give an intelligible account of this type of metaphysical freedom. According to critics, it would be nothing but an ideal aspiration that is internally incoherent.

I will now present these arguments and show that they can all be given a satisfactory response.

5.3.1. Doing away with determinism

While the most common compatibilist contention is that free will is not made impossible by the truth of determinism, a traditional argument put forward by David Hume and restated by many in the twentieth century claims that determinism is actually a most fundamental *requirement* for free will. In his influential 1943 article, "Free will as involving determination and inconceivable without it", R.E. Hobart argues that, even though our phenomenology grants us alternative possibilities ("I could have willed otherwise", in the

[427] Hard Incompatibilism is a phrase coined by Derk Pereboom (2005) to name the position according to which, just as determinism renders free will impossible, so does indeterminism.

sense that I as a person could have produced some other volition, had I so wanted[428]), it is incoherent to think that our motives *incline us without necessitating*[429], if this means that there would be an undetermined "interposition of the self" which would supersede the strength of our reasons and decide autonomously.

> "*In proportion* as an act of volition starts of itself without a cause it is exactly, so far as the freedom of the individual is concerned, as if it had been thrown into his mind from without – "suggested" to him – by a freakish demon. (…) *In proportion* as it is undetermined, it is just as if [such an agent's] legs should suddenly spring up and carry him off where he did not prefer to go. Far from constituting freedom, that would mean, in the exact measure in which it took place, the loss of freedom."[430]

According to this argument, an action is free insofar as it is controlled by the agent, and that control is rendered impossible by the presence of indeterminism which introduces chance into the causal etiology of the agent's behavior. If the agent's motives are not sufficient to ensure her decision and her action, then it is as though her intention to act (and hence her action) was not really hers, but "had been thrown into her mind from without".

Hobart is actually following a long tradition that was brilliantly defended by Hume both in his *Treatise of human nature* and in *An inquiry concerning human understanding*. Causation in nature works deterministically, in the sense that, from what we can tell after repeated observations of one event following another, there is a necessary[431] connection

[428] Hobart, R.E. (1934), p.7. He assumes a conditional analysis of "could" (see section 5.2.1), and only in that sense does he consider that an agent has the power to act and will otherwise.

[429] As in Leibniz's often quoted phrase.

[430] Hobart, R.E. (1934), p.7.

[431] Note that Hume's ground-breaking definition of causation as "constant conjunction" did not prevent him from considering causal connections as *necessary*, both in nature and in human psychology, in the following sense: "Necessity may be defined in two ways, conformable to the two definitions of cause, of which it makes an essential part. It consists either in the constant conjunction of like objects, or in the inference of the understanding from one object to another". It is in this empirical sense of inductively learnt regularity that necessity and causation can be considered to be common features of the physical and human world: "Now necessity, in both these senses (…) has universally, though tacitly, in the schools, in the pulpit, and in common life, been allowed to belong to the will of man; and no one has ever pretended to deny that we can draw inferences concerning human actions, and that those inferences are founded in the

(albeit not logically necessary) between physical causes (like the interplay of material forces) and physical effects. The same happens in the human realm: people's (re)actions are often predictable given their character and circumstances, and when they are not, that is because of our epistemic difficulties that prevent us from identifying all the causes involved. Rather than being problematic for our responsibility ascriptions, deterministic causation constitutes their fundamental precondition.

> "Where [actions] proceed not from some *cause* in the character and dispositions of the persons who performed them, they can neither redound to his honor, if good, nor infamy, if evil. The actions themselves may be blamable; they may be contrary to all the rules of morality and religion: but the person is not answerable for them (…). According to this principle, therefore, which denies *necessity, and consequently causes*, a man is as pure and untainted after having committed the most horrid crime as at the first moment of his birth, nor is his character any way concerned in his actions; since they are not derived from it, and the wickedness of the one can never be used as a proof of the depravity of the other."[432]

Insofar as a person's action is brought about by her character (motives, inclinations), together with the specifications of the circumstances, that person will be accountable for that action. If, on the contrary, an element of contingency should enter the picture (which, as a matter of fact, was considered by Hume to be impossible in our world), then the action would become an uncaused event – something that nothing, much less a free and responsible agent - contributed to bring about. And, he said, this is absurd.

Like Hobart's, Hume's reasoning too is grounded on the assumption that causation equals necessitation, which leads to his very influential idea that either an event is determined or it is purely random. David Hodgson, a contemporary libertarian, calls this (false) dilemma "Hume's mistake"[433] and shows how it is at the basis of many classical

experienced union of like actions, with like motives, inclinations, and circumstances." [Hume, D. (1748), p.97].

[432] Hume, D. (1748), p.98 (second emphasis added).

[433] Hodgson, D. (1999).

compatibilist arguments[434] and of some apparently harsh accusations against libertarianism, like Galen Strawson's following:

> "If the agent is to be truly self-determining in action this cannot be because it has any *further* desires or principles of choice governing the decisions about how to act that it makes in the light of its *initial* desires or principles of choice. For it could not be truly self-determining with respect to these further desires or principles of choice either, any more than it could be self-determining with respect to its initial desires or principles of choice. (…) But if it does not have any such desires or principles of choice governing what decisions it makes in the light of its initial reasons for action, then the decisions it makes are rationally speaking random: they are made by an agent-self that is, in its role as decision-maker, entirely non-rational in the present vital sense of 'rational' – it is reasonless, lacking any principles of choice or decision."[435].

In this quote, Strawson is presenting his version of a very serious objection against non-skeptical forms of incompatibilism, the "luck objection" (which originates in Hume-Hobart's argument), that I will address in full detail in the following subsection. For now, what I wish to unveil is the assumption that lies beneath his lines: that the only possible causal link between the agent's reasons and her decision is a deterministic one, in the absence of which choice would become unreasonable. Strawson's argument that self-determination is incoherent because it would either imply an infinite regress (of ever further motivations) or be based on purely random swerves, is grounded on the unargued premise that rationality either works as a deterministic algorithm (as if it were a matter of calculating the vector sum of all the Newtonian mechanical forces in the agent's head), or it does not work at all.

But Strawson is wrong. First, because causation can be probabilistic and, second, because human rationality need not be susceptible to be captured by deterministic rules, in order for it to retain its reasonableness.

[434] Like Alfred Ayer's (1956) or J.J.C. Smart's (1961).

[435] Strawson, G. (1986), *Freedom and belief*, cit. in Hodgson, D. (1999), p.204.

Today it is widely understood that causation is not to be identified with, nor does it entail, necessitation. One can conceive the causal relation as production[436], counterfactual dependence[437] or interpret it in some other way still, but in all these accounts there is nothing contradictory in imagining that different possible events can be alternatively derived from a single cause. Especially given the incredible success of quantum mechanics, whose standard interpretation is indeterministic, considerable work has been made in twentieth-century philosophy of science aiming at developing an account of causation that includes this possibility. It can be read as a case of *causation of probability* (the first event A is the cause of there being a 0.3 probability that a second event B will happen, a 0.6 probability that an alternative event C will happen instead, and 0.1 probability that nothing happens) or *probability of causation*[438] (in the present circumstances, there is a 0.3 probability that A will cause B, a 0.6 probability that A will cause C, and 0.1 probability that A causes nothing at all). Under both these interpretations of the probabilistic connection between A and its alternative effects, A would unquestionably be the cause of B or C, albeit an indeterministic one.

Strawson's unargued premise lies also on a conception of human rationality as an algorithmic function in which certain inputs lawfully lead to certain outputs. However, that does not have to be the default perspective and, as a matter of fact, our personal experience gives us evidence of the opposite: we often have equally persuasive reasons for making contradictory decisions and we feel that either one of those alternatives is rationally defendable. When we eventually decide, we can properly explain our choice based on the reasons we had, even though those reasons were insufficient to ensure that we would decide the way we in fact did. So, choice can be said to constitute precisely a capacity that superior conscious animals have for qualitatively resolving non-conclusive

[436] Anscombe, G.E.M. (1971).

[437] Lewis, D. (1973).

[438] I am borrowing these two contrastive formulations from Timothy O'Connor and Jonathan Jacobs' 2012 article, in which the authors defend that only the "probability of causation" perspective is adequate, in the context of their neo-aristotelian metaphysics.

reasons[439]. In his last book *Rationality + Consciousness = Free Will*, David Hodgson developed this idea at length:

> "Most human reasoning is *not* overtly algorithmic: it does not overtly proceed precisely as determined by rules of logic and/or probability and/or mathematics, or any other rules that could be incorporated into a computer program. When we are trying to make a reasonable decision as to what to believe or what to do, very often the reasons we see for and against alternative beliefs or actions are *inconclusive*, and there is an apparent *gap* between reasons on the one hand and decisions about what to believe and what to do on the other."[440]

Hodgson argues that this non-algorithmic reasoning, which he calls "plausible reasoning", is at the basis not only of most of our everyday activities but also on the basis of science. Criteria like simplicity, explanatory content and coherence with other theories, which are the subject of *ceteris paribus* rules, are crucial in fundamental processes in science such as the selection of which unrefuted hypotheses should be provisionally accepted (given the problem of the underdetermination of theory by data[441]) or the judgement of how two things are sufficiently similar for an analogy between them to be adequate. Even though plausible reasoning cannot be fully formalized, it is certainly not random as it has provided us with abundant and reliable knowledge about the world. Most importantly, Hodgson argues, plausible reasoning is what "enables us to reach decisions that resolve conflicting reasons, which are of different types and cannot be explicitly compared on a common scale"[442].

I cannot develop this argument any further here. My aim is not to tackle the complex topic of human rationality. What I meant to stress with the help of Hodgson's concept of plausible reasoning is that most of the time in our lives, we human beings have to make fallible judgements about what to believe or what to do and most often than not we are successful in this endeavor. Our decisions are intelligible to others despite their fallibility

[439] Cf. Hodgson, D. (1991, 1999, 2012).
[440] Hodgson, D. (2012), p.110.
[441] Cf. Goodman's "new riddle of induction" in Goodman, N. (1965), ch.3.
[442] Hodgson, D. (2012)

and unpredictability, and we recognize that many of the problems we face in our daily lives have more than one reasonable solution. We should beware, then, of thinking about psychology in Newtonian terms, as if our desires could be assimilated to classical forces. An argument such as Strawson's, according to which the "putative, freedom-creating power of *partially* reason-independent decision becomes a some entirely non-rational (reasons-independent) flip-flop of the soul"[443], is likely to be based on such a mechanistic assumption.

However, the most important objection to libertarianism is yet to come. Even if we accept that an agent can produce one or more different decisions given the exact same circumstances and laws of nature, does this imply that the Hume-Hobart objection fails? Not yet, for at least in Hobart's case, the argument was not only that indeterministic causation of an action was most likely false but also that, if it were true, it would reduce (rather than enhance) the agent's control. But this brings us fully to the luck problem, one which has been abundantly discussed in the recent literature on free will and constitutes arguably the greatest threat against libertarianism.

5.3.2. The luck problem

The luck objection against libertarianism has assumed several different formulations in the literature. The two main lines have to do with explanation and control. The problem underlying both of them can be appropriately described by means of the following example by van Inwagen:

> "Let us consider the case of a hardened thief who, as our story begins, is in the act of lifting the lid of the poor-box in a little country church. He sneers and curses when he sees what a pathetically small sum it contains. Still, business is business: he reaches for the money. Suddenly there flashes before his mind's eye a picture of the face of his dying mother and he remembers the promise he made to her by her deathbed always to be honest and upright. This is not the first occasion on which he has had such a

[443] Strawson, G., *op.cit*.

> vision while performing some mean act of theft, but he has always disregarded it. This time, however, he does not disregard it. Instead, he thinks the matter over carefully and decides not to take the money. Acting on this decision, he leaves the church empty-handed."[444]

If we assume that the thief's decision was undetermined, then there can be an alternative world (a hypothetical world that is as similar to the actual world as possible) in which, in spite of the identical conditions preceding the decision, he decides to steal instead. However, this possibility results in a problem with two horns:

1) If all the reasons and considerations that preceded his decision in the alternative world were exactly the same as they are in the actual world, what is it that explains his deciding to steal rather than to refrain, and vice-versa?
2) If all the conditions were already settled and still the future was open, then how can we say that the thief is in control of his decision?

If the difference between worlds cannot be adequately explained and the decision cannot be controlled by the thief[445], then one can say that the fact that he refrained in the actual world is not a doing he should be praised for. It was just a matter of moral luck.

Even though these two horns are intertwined (for instance, the putative lack of explanation is said to be revealing of the putative lack of control) I will address them separately for the sake of clarity.

5.3.2.1. The lack of contrastive explanation

Why should the lack of an adequate explanation be a problem? What Alfred Mele (one of the main proponents of the explanatory formulation of the luck problem) is worried about is that if there is nothing that accounts for the difference between the actual world in

[444] van Inwagen, P. (1983), pp.127-8.

[445] Cf. Mele, A. (2006), chapters 1 and 3, for an often cited presentation of the luck problem in terms of cross-world differences.

which the thief decided to refrain and an alternative world in which he decided to steal, then this difference is just a matter of luck. And luck is incompatible with free will, which leads us to conclude that indeterminism precludes the thief from being free, rather than allowing him to be so.

There are many assumptions in between the lines of this formulation. First, what counts as an adequate explanation of an indeterministically caused action? Second, what makes one infer that the lack of explanation for an action implies that it was just a matter of luck? Third, why should we assume that luck undermines free will? I will now address these assumptions one by one.

a) What counts as an adequate explanation of an indeterministically caused action?

In the context of the free will debate, it has sometimes been assumed that any non-deterministically caused action would be unintelligible (Ayer, for example, said that if a choice is explicable, then "we are led back to determinism"[446]). Nevertheless, most contemporary authors agree that for any action made for reasons, determined or undetermined as it may be, we can have a satisfactory and reasonable explanation that refers to the reasons on which the agent acted. For instance, regarding the above mentioned case, Mele would not question the fact that we can cite the memory of the promise the thief made to his mother as the reason that explains why he refrained. However, since that memory was present in the alternative scenario as well, it cannot *contrastively* explain the action, which is to say it cannot explain the difference between the two worlds from the moment of the decision on. In fact, the explanation that many consider to be lacking in cases of undetermined actions is contrastive explanation: the explanation of why the thief decided to refrain *rather* than to steal[447].

There are two complementary answers to this problem. One is to question the assumption that a contrastive explanation cannot be given in cases of undetermined decisions and

[446] Ayer, A.J. (1954) cit. in Clarke, R. (2003), p.31.

[447] Cf. Sorabji, R. (1980), Russell, P. (1984), Nagel, T. (1986).

actions. The other is to question that that form of explanation is the only adequate explanation in these cases[448].

First of all, we must acknowledge that there are many different kinds of situations in which an agent performs a free action and that, depending on the peculiarities of each case, more or less satisfying explanations will be available.

On an agent-causal account such as mine, actions are caused, but not necessitated, by the agent. The agent's reasons (her beliefs and desires) influence the probabilities of each outcome, insofar as the agent has a greater tendency to act on a stronger reason than on a weaker one. When the agent acts on her stronger reasons, contrastive explanation is not a problem. Peter Lipton's account of contrastive explanation[449] shows why this is so. According to him, one can have a good contrastive explanation for an indeterministically caused event if there is an explanatorily relevant factor that raised the probability of the actual event happening and which was in a certain relation to that event, without there being any corresponding factor in the causal history of the alternative event. Randolph Clarke presents this with the help of an example he borrowed from Humphreys:

> "The bubonic plague bacillus (*Yersinia pestis*) will, if left to develop unchecked in a human, produce death in between 50 percent to 90 percent of cases. It is treatable with antibiotics such as tetracycline, which reduce the chance of mortality to between 5 percent and 10 percent."[450]

In this case, despite the less-than-unity probabilities of both events (surviving and dying), the antibiotic can contrastively explain why a certain patient died and another one lived (or why a certain patient lived in the real world while his counterfactual twin died in an alternative world) by referring to the fact that the survivor took the antibiotic while the untreated patient did not. There is no corresponding event in the causal history of the latter patient that raised the likelihood of his death to close to 1.

---

[448] Randolph Clarke (2003), Timothy O'Connor (2000) and Christopher Franklin (2011a, forthcoming a) are amongst the authors that in the past years have strived (in my view, successfully) to make their case against the luck problem based on both these strategies.

[449] Lipton, P. (1990, 1991, 1993). His account stems from Mill's "Method of Difference".

[450] Humphreys, P. [(1989), p.100] cit. in Clarke, R. (2003), p.41.

Assuming that rational explanation is a species of causal explanation, Lipton's model applies straightforwardly to cases of reasons-based decisions: if there is a strong reason or group of reasons that makes decision A preferable to decision B (but still not inexorable), then if the agent acts on those reasons his action will be contrastively explainable by citing them[451].

Not all decisions are like this, however. There is always the possibility that the agent will do something that was very unlikely. In other cases still, she will find herself before tied-for-best options, both sustained by equally good and motivationally strong reasons: reasons for A on which the agent will act in case he decides to do A, and reasons for B on which the agent will act in case he decides to do B. In situations such as these, when the agent does not act on her better and stronger reasons, the action can still be rationally explained by citing those reasons that *did* support it – but there is no contrastive explanation available.

This is not a problem, however, since a noncontrastive explanation is not incomplete. It is informative, rational and causal and if it fails to show how a certain decision was inevitable or more probable than its alternative, that is simply because it was not and there is no more information to be had. Clarke believes the frequent misjudgment about the adequacy of a noncontrastive explanation that cites all the relevant information about the causal history of the *explanandum* has to do with the false assumption that a full explanation is one that tells us why the event had to happen, which is obviously fallacious in the case of indeterministic events:

> "It might be accepted that nondeterministically caused events can be explained but objected that they cannot be completely or fully or adequately explained because it cannot be explained why they had to happen. But the question why such an event had to happen carries a false presupposition; the event did not have to happen. And it is

[451] This is no novelty. On Hempel's model of scientific explanation, the possibility of an inductive-statistical explanation relied also on the degree of inductive support that the *explanans* gave to the *explanandum*. That degree could not be numerically fixed but it was required that it be close to one.

> no incompleteness or inadequacy in an explanation that it fails to answer to a false presupposition of an explanatory question."[452]

Another very important consideration is that the criteria for the adequacy of an explanation depend on the knowledge we have prior to the presentation of the data cited in the explanation. As Christopher Hitchcock has shown[453], we can have technically correct contrastive explanations which are nevertheless pragmatically inadequate (and thus fail to explain the contrast between the actual event and its alternative) because they cite information that was already presupposed. The exact same datum may be relevant or not, just as the same fact may be explainable or not, depending on epistemic conditions. In the aforementioned case of the bubonic plague patient who was treated with the appropriate antibiotic, referring to that treatment may be perceived as inadequate, say, by the family of a roommate of his, whom, by hypothesis, was in the same condition as him but died despite having taken the antibiotic. In such a case, that family would already know that the surviving patient had taken the antibiotic and that it had raised enormously his probabilities of healing, and what they would want to know was some further cause that made it so that, given the same probabilities, he survived while their loved one did not. But if we assume that, given the condition of both having been appropriately treated, the event of dying or surviving is truly indeterministic, then there is no more information to be had. The family of the patient who died already knows everything that can be cited in an explanation, and thus any explanation will be unsatisfactory for them.

Once again, this reasoning applies to the explanation of actions as well. If we imagine God, an omniscient being, as being in possession of every single detail of a free agent's mental life, then we may say that no explanation of why the agent chooses to act as she does can sound pragmatically adequate to His ears. However, God can give a contrastive explanation of the agent's most probable actions to anyone who does not possess all the relevant information in advance, as well as a reasonable explanation of those actions that

[452] Clarke, R. (2003), p.36. On the idea that contrastive explanations can sometimes be given for undetermined actions, but they are not necessary for a rational and adequate explanation of action, see also O'Connor, T. (2000), section 5.3.

[453] Hitchcock, C. (1999).

result from choices that were unlikely. Both explanations will allow one to claim that the agent's actions, having been done for reasons, are not random nor arbitrary.

b) What justifies the inference that the lack of explanation for an action implies that it was just a matter of luck?

This question requires one to make clear what luck is supposed to mean. In his aforementioned argument, Mele used a definition of luck that is actually question-begging and cannot serve as a justification for the inference from the lack of contrastive explanation to the conclusion that the cross-world difference is just a matter of luck:

> "[I]f the question why an agent exercised his agent-causal power at t in deciding to A rather than exercising it at t in any of the alternative ways he does in other possible worlds with the same past and laws of nature is, in principle, unanswerable (…) and his exercising it at t in so deciding has an effect on how his life goes, I count that as luck for the agent.[454]"

Mele defines luck as the absence of contrastive explanation, which is plainly circular. But maybe we could use a different definition. Neil Levy also points out that the luck objection is crucial, but he grounds his definition of luck differently. According to him, luck is the agent's lack of control over the occurrence of a rare event of personal significance to her[455]. Could the definition of luck as lack of control over an improbable event fare better in sustaining the implication from the lack of contrastive explanation to the presence of luck as the determinant of the undetermined event? I do not think so, since issues

[454] Mele, A. (2006), p.70.

[455] Neil Levy considers there are two types of luck, both grounded on the lack of control the agent has for an improbable event that is significant for her: "An event or state of affairs occurring in the actual world is *chancy lucky* for an agent if (i) that event or state of affairs is significant for that agent; (ii) the agent lacks direct control over that event or state of affairs, and (iii) that event or state of affairs fails to occur in many nearby worlds (…). An event or state of affairs occurring in the actual world that affects an agent's psychological traits or dispositions is *non-chancy lucky* for an agent if (i) that event or state of affairs is significant for that agent; (ii) the agent lacks direct control over that event or state of affairs; (iii) events or states of affairs of that kind vary across the relevant reference group, and (iv) in a large enough proportion of cases that event or state of affairs fails to occur or be instantiated in the reference group in the way in which it occurred or was instantiated in the actual case" [Levy (2011) p.36, emphasis added].

regarding explanation are different from issues regarding control. Again Randolph Clarke helps us clarify the matter:

> "A good explanation answers well a question that we have, or one that we could sensibly ask about an occurrence. In some cases, there may not be available an explanation that answers well a quite sensible question, but such unavailability need not correspond to any lack of active control. Further, if substance causation is possible, then some causes may contribute crucially to the exercise of active control and yet citing them, while explanatory, may not yield the sort of explanation in which we are typically interested. And finally, an explanation may succeed without citing phenomena that constitute an exercise of that variety of active control that is required for acting freely, for such phenomena may not be present. Hence, we may have adequate rational explanations where free will is lacking."[456]

Hence, we can sometimes have a good explanation for facts which are beyond any agential control, as well as no explanation for facts that are in fact controlled by an agent.

Also, as Hitchcock's account showed us above, lack of contrastive explanation depends on epistemic and pragmatic conditions. Matters of free will and control are matters of what is the case in the world (what causes what and who can ensure that something will occur rather than something else), and thus they cannot be subject to such factors as the previous knowledge one has about an event.

c) Why should we assume that luck undermines free will?

The last unfounded inference in Mele's argument is that if an action is a matter of luck, then it cannot be free. This seems highly intuitive if we, like Levy, take luck to mean lack of control for an improbable event. If the first time I play basketball I score a three-point field goal, I will consider that to be due to the so called "beginner's luck" and certainly not to my skills. The reason for this judgment is that I am perfectly aware that if I try to score

[456] Clarke, R. (2003), p.32.

again, I am very likely to miss, so much so if I try to many times. I am not in control of my performance, hence my scoring happened in spite of me.

However, we have seen above that such a definition of luck is not suited for the entailment Male wants to make. For Mele's argument to work, we must assume luck to be just another word for the absence of factors that account for the cross-world difference, but then the conclusion that luck undermines freedom will not follow. From the fact that there is no prior difference between the causal histories of the two alternative worlds that might account for the two different decisions, one cannot infer that the agent was not free to act as she did. Incompatibilist freedom means being able to choose differently under the same circumstances, and the lack of explanation cannot undermine this ability without undermining some further feature that proves essential for it.

The idea that the lack of contrastive explanation reveals lack of control is one way to make this move, but, as we have seen, a fallacious one. We can have control without explanation and explanation without control. Another way to link the lack of contrastive explanation to free will is to say that the former reveals that the agent's reasons were insufficient to account for her decision, which means that the action was not *made for reasons*. If nondeterministic causation would preclude acting for reasons, that would certainly undermine human free will[457] and the possibility of action, for that matter. However, this strategy relies on the assumption that the availability of a reasons explanation of an action depends of the availability of a contrastive explanation of it, which is false, as we also have seen above. Any intentional action is made for reasons (albeit often unconscious ones) and we can adequately explain it by citing those reasons. If we cannot access them due to our epistemic limitations, that does not entail that the action was not motivated by a belief-desire pair and caused by the agent for that reason.

[457] Cf. Levy, N. (2011).

5.3.2.2. The problem of diminished control

Let us now go back to van Inwagen's case of the thief who, upon deliberation, refrains from stealing, in order to tackle the second horn of the luck problem. Does the fact that he could have chosen otherwise under those exact same circumstances diminish his control over his decision, with respect to an alternative scenario in which his action was causally determined?

No, it does not, for the thief's decision was a causing of his. In deciding, he exercised his downward causal power over his brain and body and made sure that his action was to refrain. The fact that he could have decided differently means only that he could have exercised this same power in another direction. And if the universe were deterministic, he would not have been able to exercise that power at all, since the complete state of his brain and body would have been entailed by their previous state, immediately preceding the decision. Therefore, not only is indeterminism not threatening for control, as it is an absolutely necessary condition in order to avoid causal redundancy[458].

As we have seen in section 5.3.1, indeterministic causation does not amount to any causal gap in the unfolding of events. Therefore, the fact that there are two or more possible sequences does not diminish the agent's active role in the actual sequence that *does* happen. In the words of Randolph Clarke:

> "The nondeterministic nature of the causation (or of the governing causal law) is just a matter of similar agents' elsewhere behaving differently, something that does not imply any weakening of any token relation between token mental events involving this agent and her token action."[459]

Clarke[460] suggests we can see this clearly if we imagine two identical agents in two parallel worlds, a deterministic and an indeterministic one, that share the same physical history (every single event that has taken place in their bodies and surroundings) up until the

[458] I will develop this idea in section 5.3.3. on the problem of enhanced control.
[459] Clarke, R. (2003), p.73.
[460] Clarke, R. (1995; 2003), ch.5

moment of decision. In analyzing their decision process, one will endorse a particular account of causation (a humean account or a realist account, for instance), and logically one will have to keep that account as the suitable description of the causal process that will take place in each of the two worlds. This means that either there is no real connection between events in any of the worlds (humean account of causation), or there is an effective link between them (realist account). What changes between worlds is the nature of the higher-level laws that regulate (in a necessitarian versus probabilistic manner) the causal relation. If we are imagining the same decision in both worlds, we will find one cause (the agent and/or her mental events, according to the libertarian account at stake) that will produce the same effect (a certain decision), via the same type of causal relation, with the only difference that, in the indeterministic scenario, some other decision would have been possible as well.

But maybe this response can still be countered by further arguments. Let us now address two influential examples that van Inwagen has used in favor of the thesis that indeterminism is control undermining. The first one is known as the "rollback argument" and it features Alice, who is having to choose between lying and telling the truth. Since she is a libertarian agent and this will be a free choice, both alternatives are compatible with her past and the laws of nature. Then van Inwagen suggests we do a thought experiment:

> "Now suppose that immediately after Alice told the truth, (…) God a thousand times caused the universe to revert to exactly the state it was in at t_1 (…). [W]e cannot say what would have happened, but we can say what would probably have happened: sometimes Alice would have lied and sometimes she would have told the truth. As the number of "replays" increases, we observers shall – almost certainly – observe the ratio of the outcome "truth" to the outcome "lie" settling down to, converging on, some value. (…) If we have watched seven hundred and twenty-six replays, we shall be faced with the inescapable impression that what happens in the seven-hundred-and-twenty-seventh replay will be due simply to chance."[461]

[461] Van Inwagen, P. (2000), pp.14-15.

Van Inwagen's conclusion is that Alice's action, just like any undetermined action, is just a matter of chance (he never uses the word luck, even though his example is considered to be one of the formulations of the luck argument). Given the fact that there is an objective chance she might lie (a chance the value of which we can define with increasing precision as the replays take place), we should consider ourselves fortunate if she were to tell us the truth. Her action is a lucky event, and this prevents it from being free.

> "If she was faced with telling the truth and lying, and it was a mere matter of chance which of these things she did, how can we say that – and this is essential to the act's being free – she was *able* to tell the truth and *able* to lie? How could anyone be able to determine the outcome of a process whose outcome is a matter of objective, ground-floor chance?"[462]

In order to show that objective chance prevents control, van Inwagen adds another example still, which I will adapt to his rollback case, for the sake of simplicity. Imagine that I am a libertarian agent and that, instead of having to choose between being truthful or false like Alice, I am asked to promise that I will not reveal a secret. There are good reasons both for telling the secret and for keeping it. Just like in the previous case, there would be a less than unity chance for each alternative: a 0.43 probability of my telling the secret, against a 0.57 chance of my keeping it, to be more precise. The question is: if I would come to know the value of the probability of my failing to keep my promise, would I dare to make that promise nonetheless? Van Inwagen says I should not:

> "Am I in a position to promise you that I will keep silent? – knowing, as I do, that if there were a million perfect duplicates of me, each placed in a perfect duplicate of my present situation, forty-three percent of them would tell all and fifty-seven percent of them would hold their tongues? I do not see how, in good conscience, I could make this promise. I do not see how I could be in a position to make it. But if I believe that I am able to keep silent, I should, it would seem, regard myself as being in a position to make this promise. What more do I need to regard myself as being in a position to

[462] *Idem*, pp.15-16.

> promise to do X than a belief that I am *able* to do X? Therefore, in this situation, I should not regard myself as being able to keep silent."[463]

Are these two cases knock-down arguments against libertarianism? Do they make my response to the control formulation of the luck argument ineffective? I will now move on to analyze them. Just like in the explanatory formulation presented in the previous section, here too we find different steps that need to be addressed: first, the idea that indeterminism implies objective probability; second, that objective probability implies lack of control. I will contend that they are both unwarranted assumptions.

a) Does indeterminism imply objective probability?

The first assumption leads van Inwagen to take for granted that, in a hypothetical situation of numerous replays of the same decision under identical conditions, we would "most certainly" find that the ratio between the two alternatives converges to a real number. Lara Buchak makes a very strong case against this supposition. Van Inwagen's inference is based on the laws of large numbers. This is very clear in an endnote where he adds that "as the number of replays increases, the probability of 'no convergence' tends to 0"[464]. However, Alice's case is very different from a coin flip case in which, given that we know there is an objective probability associated with each toss, we can reasonably expect there to be convergence: as the number of trials increases, the ratio of each outcome to the total number of trials will reflect the objective probability of each single outcome. In Alice's case, however, the objective chance of each decision is precisely what we are trying to demonstrate! As the law of large numbers presupposes the objective probability of the outcomes, it cannot be used to prove it. In the words of Buchak:

> "The rollback argument directly begs the question of whether Alice's lying has an objective probability. Without the assumption that it does, (…) there is nothing at all to rule out, for example, the following series of choices: the first time God reruns the

[463] *Idem*, p.17. Ishtiyaque Haji's renowned "ensurance formulation" (2001, 2004) of the luck problem is very similar to the argument put forward by van Inwagen with the promise case. I consider that the responses I present here against the latter can also be used against the former.
[464] *Idem*, p.19, note 16.

> situation, Alice lies; the next 9 times, she tells the truth; the next 90 times, she lies; the next 900 times, she tells the truth; and so forth. In this example, the proportion of lies never converges (it will alternate between roughly 1/11 and 10/11, after each 10n trials). Contra van Inwagen*, there is nothing in his setup even to make this unlikely*. Unlike in the coin-flipping case, there may not be a chancy mechanism – or a mechanism that *behaves as if* it is governed by chance – grounding Alice's actions."[465]

And, if there are no objective values to be attributed to the probabilities associated with each outcome, then the promise example loses its basis as well[466].

b) Does objective probability entail the lack of control?

The main question the rollback formulation of the luck argument raises is whether, in case Alice and I do have an objective chance of doing A and an objective chance of doing B, we can say that we are able to cause A or B. Van Inwagen says we cannot and he uses the case of me being unable to assure that I will keep my promise as an argument for this. Even though both his examples seem very intuitive and strong, I believe they fail to prove his conclusion, and I will now show why this is so.

Christopher Franklin[467] notes – and I follow his lead on this – that one of the problems with van Inwagen's account is that it has the location of indeterminism wrong. He is describing a case in which I decide to keep a secret (and therefore I make a promise not to tell), but there is an indeterministic link between my intention to keep my promise now

[465] Buchak, L. (2013), p.24.

[466] In his rollback case, van Inwagen makes use of a frequentist interpretation of the concept of objective probability. The problem that Buchak's objection reveals can actually be seen as a special case of the more general problem that this interpretation faces: circularity. For example, when one defines the law of large numbers as follows: "In repeated, independent trials with the same probability p of success in each trial, the percentage of successes is increasingly likely to be close to the chance of success as the number of trials increases" [Stark, P.B., "Glossary of Statistical Terms", cit. in Buchak, L. (2013) p.23], one is assuming that p exists; however, according to the frequentist definition of probability, there is no such thing as the objective probability of a single trial. The value p that figures in the law is to be defined as the limit to which the value of the relative frequency with which each outcome occurs in an infinitely extended series of trials is supposed to converge.

[467] Cf. Franklin, C. (2011).

and my action in the future. In such a scenario, the fact that the objective probabilities are already fixed would be control-diminishing as this type of account opens up the possibility that the agent's intention might not lead to what was meant. The objective probabilities that I come to know in advance will in fact prevent me from making any promise because they reveal to me my lack of control over my future action, regardless of my present decision. In this situation, indeterminism would in fact introduce an element of risk in the causal relation that connects together my intention to act and my overt action, which would make it so that after exercising my active role in the deliberative process, I might as well just sit back and wait to see if, luckily, my intention leads to the best action or not. The outcome is still open but there is nothing left for me to do.

Real situations in which libertarian agents are implicated are not like this at all. When I decide to make a promise not to tell a secret and I make it, I am acting *now* (first action) and changing the future probability of my telling that secret (a second action). That probability had kept changing for various reasons up until now, the moment when I finally intervene in the course of events and make my final decision. If my character and background have endowed me with self-control and values such as truthfulness and loyalty, from this moment on my action will outflow from my decision with a probability close 1.

If libertarianism is to avoid the problem presented by the promise case, it has to postulate the location of indeterminism between the agent's reasons (what Franklin calls "non actional mental states"[468]) and her choice. The choice is the moment when the agent causes the action to happen: in Alice's case, it is the moment when she determines whether she will lie or tell the truth. From that moment on, her action is as determined in an indeterministic world as it would be under determinism: it will take place, unless something external to the agent's agential apparatus (her reasons and her will) prevents it (for example, in case she faints or dies, or if something calls her attention and she interrupts what she was doing, etc.). Thus located, indeterminism does not diminish the

[468] Franklin, C. (2011), p. 205.

agent's control, it gives her the possibility of exercising the rational control she is endowed with in one direction as well as in another[469].

At this point it is important to note that indeterminism is not an extra causal element in the course of events. Indeterminism is the negation of determinism: when we admit that something was "just a matter of chance", what we mean to say is that it was not inevitably determined to happen by what took place before nor by the agent-causal capacity of an agent. Indeterminism is a negative condition, not a positive intervener in the world. This is a relevant consideration because it opens room for the agent to be that positive intervener, taking advantage of the openness the absence of determinism grants her. The agent has reasons for and against each alternative and she has the ability to exercise her power to cause one of both alternatives, since indeterminism opens up the possibility for her to do so.

5.3.3. The problem of enhanced control

In the previous section we have seen that libertarianism can address the luck objection under its various formulations and I hope to have shown that, unlike what the objection seemed to stress, indeterminism *per se* does not diminish the agent's control. However, some might doubt that I have also made clear that indeterminism can enhance the agent's power to choose, relative to the power she possesses in a deterministic world. Compatibilist agents are "free" in the sense that they can, in normal circumstances, act according to their reasons. Is a libertarian agent free *also* in the deeper sense of being in charge of her intentions, and thus having genuine alternative possibilities of action?

Opinions differ, even amongst libertarians, of course. Event-causalists[470] consider that their version of action causation is sufficient to ensure that the libertarian agent possesses more control in an indeterministic rather than in a deterministic world, while agent-

[469] Cf. Franklin, C. (2013) on the most appropriate location of indeterminism.
[470] Cf. Mele, A. (2006), Franklin, C. (2011, 2014b).

causalists[471] disagree. As one might expect, given the account that I have developed in the previous chapters, especially in the first one, my view is that event-causalist accounts do fall prey to the luck objection, because of the associated problem of the disappearance of the agent, which I have already presented in section 2.6. In her 2010 article "Why agent-caused actions are not lucky", Megan Griffith shows how the two problems are associated and how they make it so that only an agent-causal agent possesses the type of control required for free will. According to her, the unavailability of a contrastive explanation for cross-world differences is not a problem for libertarianism but reveals the real threat facing it: the lack of control over *which* decision is made. And that problem can only be overcome by an agent-causal account, in which the agent's prior mental states do not exhaust the action's total cause.

> "In both the event-causal and the agent-causal cases, the unavailable explanation corresponds to the causal openness of the decision, given the prior states of the agent. But this does not indicate lack of control for the agent-causal case, since the prior states of the agent are not the whole story. It does indicate a lack of control for the event-causal case because the agent has no further causal involvement in the action (…) As such, it just happens to her that the decision is to A rather than to B because she is missing the power to determine the decision."[472]

The opportunity to choose that is granted to the agent by indeterminism is differently explored in the event-causal and the agent-causal scenarios. In the former case, the agent controls the choice she makes insofar as she is the one who produces it, but she does not possess what I shall call *contrastive control* over which alternative becomes actual. In turn, in the latter scenario, the opportunity is explored by the agent-causal agent by exercising her ability to choose.

Interestingly, van Inwagen's primary intention in presenting the above two examples (the rollback and the promise case) was to show that agent-causation could not save the

[471] Cf. O'Connor, T. (2000), Griffin, M. (2010). Also the agnostic Randolph Clarke (2003) and the skeptic Derk Pereboom (2004, 2014) claim that only agent-causal libertarianism can enhance the agent's control.

[472] Griffith, M. (2010), p.51.

agent's control, given objective chance. After presenting the promise example, he imagined a situation in which God would reveal to him that he would agent-cause his action, whichever alternative would eventually occur. And so he asks:

> "Why should [the revelation that I am the agent-cause of events in my brain that will result in the bodily movements that constitute my act] lead me to conclude that I am in a position to promise to keep silent and therefore that I am able to keep silent? Its content simply does not seem to be relevant to the above argument for the conclusion that it is false that I am able to keep silent."[473]

Remember Franklin's response to van Inwagen's first presentation of the promise example: given the location of indeterminism between the coming to mind of the reasons to act and the agent's decision, we can reasonably consider that the openness of different alternatives does endow the agent with the possibility to be the determining cause of the occurrence of the action of keeping the promise *as well as* of the action of breaking it. But I would add: in order for that possibility to become an actual ability, the agent as such must be the cause of the action – apart from her reasons, which have all been previously given at the moment of decision. If the agent is reduced to (some of) her reasons, then there is no way she can choose which of the alternatives will become actual. I believe, in fact, that in the event-causal context Franklin should not refer to the act of making a decision with the term "choice". Deciding, in his sense, is *causing*, but not *choosing*.

The incompatibilist event-causal account endows the agent with the possibility of behaving differently for different reasons and, as we have seen, this does not represent a reduction of the type of control that her compatibilist counterpart possesses. However, given the disappearing agent objection, that control is not sufficient. The agent-involving appropriate mental states cause one of the alternatives, true, but it is not up to the agent which of them they cause. Instead, I believe that in order for the agent to have "freedom-

[473] Van Inwagen, P. (2000), p.18.

level"[474] control, she needs to be the author of her action – its "source"[475], its "originator"[476]. In Clarke's words:

> "The presence in the indeterministic world of a *chance* that is absent from the deterministic world does not by itself give the indeterministic agent a further *ability*. To have a better variety of control over her behavior than that exercised by her deterministic counterpart, an agent in an indeterministic world must have a further power to determine which of the actions, each of which she might perform with control, she will actually perform."[477]

The type of control that agent-causation grants the agent in an indeterministic context is contrastive control: the control not only over "A" and over "B" but also over "whether A or B". That, I contend, is the enhanced control the agent needs in order to perform actions that are both appropriately authored (as any agent-caused determined action is) and free. And as we have seen in the previous chapters, this type of control is what the agent's emergent conscious self can exercise with its downward causal power over the brain that brings it about.

[474] Clarke, R. (2003) p.67.
[475] Pereboom, D. (2003).
[476] Clarke, R. (2003), p.160.
[477] *Idem*, p.134.

6. CONCLUDING REMARKS

The itinerary we have undertaken in this dissertation has reached its endpoint. I presented a new argument for incompatibilism and showed how we have no reasons for endorsing a nihilist position. Agent-causal libertarianism is the only way to vindicate our experience as agents and it comes at a much lower cost than is usually assumed.

Some of the theses that sign this itinerary have been presented along the way, others still need to be stated explicitly. Let me now sum them all up.

First, regarding what an action is. Being caused by some of the agent's mental states and events is not a sufficient condition for a behavior to be considered to be an action. For an action to be such, the agent must be its ultimate substance-cause. What I mean by this is that the agent's self makes a choice about what to do, thus producing an intention to act. The agent's action is the composite event of the agent's producing an intention to act which then causes the bodily movement to occur.

Second, regarding the agent as cause. The agent's intentionally causing an action amounts to more than its parts (e.g. neurons, muscles) being involved in natural causal sequences (e.g. synapses, movement), which sometimes take place in the absence of the phenomenal experience of agency. The distinctiveness of agent-causing requires a self as an irreducible entity with non-derivative causal powers. The agent's *conscious* self, which emerges naturally from a sufficiently complex brain and is the bearer of the agent's conscious properties, is such an entity.

Third, regarding the conditions for agent-causation. Given the supervenience of mental states and events on physical states and events, the agent's self can only produce an intention to act by downwardly causing the occurrence of certain neural events which, in turn, will cause some bodily movements. In order to avoid causal redundancy and epiphenomenalism, the possibility of downward causation requires the break of both

bottom-level determinism and causal closure of the physical. Hence, agent-causation entails physical indeterminism.

Fourth, regarding freedom of the will. For the agent to be the determiner of an action, her will, i.e., her power to make decisions or to form intentions to act, cannot itself be determined to choose in a certain way. Nevertheless, the agent's will is partially constrained by her beliefs and desires which, in turn, are a consequence of her biological traits and environmental factors. That influence does not determine the agent's choice; it structures the probabilities of the various alternatives she can choose from and then leaves room for her downward causal power to be exercised. This is made possible by neuronal indeterminism whereby the state of the agent's brain at a certain instant is compatible with more than one immediate after-states.

Fifth, regarding action explanation. At the mental level, neuronal indeterminism corresponds to the agent's having different inconclusive reasons in favor of different options. Given this openness of possible futures, the agent's self can intervene in the world and cause one state to happen, rather than another, without thus contradicting any laws of nature. Still, her action is influenced, though underdetermined, by the physical constrictions and psychological reasons that raised its likelihood and which ought to figure in its adequate rational explanation.

Sixth, regarding the possibility of unfree actions. Involuntary action is a contradiction in terms. All actions are agent-caused and the agent's self cannot cause anything that already has a sufficient cause that the agent cannot control. Hence, all actions stem from a choice the agent makes from amongst open alternatives. I endorse an incompatibilist view of action (there can be no deterministically caused actions) that is also an incompatibilist view of free will (there can be no free will in a deterministic world). All actions are free in a libertarian sense.

Seventh, regarding the role of consciousness. Even though all actions are free, not all of them are *directly* free, insofar as people sometimes act on standing intentions of which they are not presently aware. When that happens, those behaviors can nevertheless count as *indirectly* free actions, since they are the expression of preferences and

tendencies to act that were formed through previous conscious and deliberate choices of which the agent was the direct cause.

Eighth, regarding the scope of agency. Despite all the physical, environmental and psychological constraints that limit the range of possibilities an agent has at the moment of choice, very seldom are the possibilities reduced to one. Most of the time, our supposed actions are not determined behaviors in disguise – they are true actions, brought about by our emergent selves who have the ability to make choices and to settle how our lives, and the portions of the world they influence, unfold.

An account such as mine is not easy to defend in the context of current analytical philosophy. It relies on metaphysical assumptions that are insusceptible of empirical control and it goes against the important requirement of parsimony. However, as Daniel Dennett once wrote, we must never mistake "a failure of imagination for an insight into necessity"[478]. Ironically, while Dennett meant this motto to be a warning against the risks of making statements about reality (such as, for instance, arguments for the in-principle impossibility of functionalizing consciousness), based only on our present partial knowledge about the world and our psychological and theoretical biases, I find it very appropriate when it comes to deflationist positions such as his.

I certainly believe philosophy must be scientifically informed, and that is why, throughout this dissertation, I had the concern to show how the positions I endorse do not contradict what science is currently committed to. This is the common ground one should assume as a fixed point. Nevertheless, the requirement for empirical plausibility must be well balanced with that of broad rational plausibility. And under this criterion, many of Dennett's (and others') positions are highly implausible, despite their metaphysical unproblematicity and scientific appeal. I believe they limit their imagination to scientific inquiry and assume science's method of posing questions to nature to be our only way of having an insight into reality.

[478] Dennett, D. (1993), p.401.

My agent-causal theory provides us with theoretical tools to account for the (auto and hetero) phenomenological experience we have as agents in the world, allowing us to justify the empirical distinctions we make between the cases in which our behavior merely outflows from our mental states and those in which we are the protagonists of the choices that sign our path. By exploring the idea that the agent-cause is the agent's conscious self, my proposal conjoins the already entangled subject matters of free will and consciousness, which are recognized as open problems both in philosophy and in science. Despite its eagerness to go beyond the limits of what can be empirically tested, my account is compatible with neuroscientific knowledge about the functioning of the brain, as well as with what physics shows to be the truth behind the myth of reductionism.

I obviously do not expect to have given a definitive answer to the philosophical problems that I am tackling here. I suspect that they have no solution. Nevertheless, I do hope to have added a new line of argument to the debate, to have justified it with honest reasoning and proven its possibility, reasonableness and even plausibility with sufficient interdisciplinary knowledge. This is how I believe there can be progress in philosophy.

BIBLIOGRAPHY

Alexander, S. (1920), *Space, Time, and Deity*, London: Macmillan.

Anderson, P. W. (1972), "More is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science", in Bedau, M.A., Humphreys, P. (2008), *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge, MA: MIT Press, pp.221-229.

Andrade e Silva, J., Lochak, G. (1988), *Quanta, Grãos e Campos*, Lisboa: Instituto das Novas Profissões.

Anscombe, G.E.M. (1971), "Causality and determination", in Sosa, E., Tooley, M., eds. (1993), *Causation*, Oxford: Oxford University Press, pp. 88-104.

Armstrong, D. M. (1997), *A World of States of Affairs,* Cambridge: Cambridge University Press.

---------- (1999), *The mind-body problem. An opinionated introduction*, Boulder: Westview Press.

Arp, R. (2008), "Emergence in Biology", *Cosmos and History: The Journal of Natural and Social Philosophy* 4: pp.260-285.

Arpaly, N., Schroeder, T. (1997), "Praise, blame and the whole self", *Philosophical Studies* 93: pp.161-188.

Austin, J.L. (1956), "Ifs and Cans", in Austin, J.L. (1979), *Philosophical Papers*, Third Edition, Oxford: Clarendon Press, pp.205-32.

Ayer, A.J. (1956), "Freedom and Necessity", in *Philosophical Essays*, London: Macmillan.

Baars, B. J. (2001), *In the Theater of Consciousness: The Workspace of the Mind*, New York: Oxford University Press.

Baker, M., Goetz, S., eds. (2011), *The Soul Hypothesis. Investigations into the Existence of the Soul*, New York: Continuum.

Balaguer, M. (2009a), *Free Will as an Open Scientific Question*, Cambridge, MA: MIT Press.

---------- (2009b), "Why There Are No Good Arguments for any Interesting Version of Determinism", *Synthese* 168: pp.1-12.

Bandyopadhyay, N.J., Paterek, T., Kaszlikowski, D. (2012), "Quantum Coherence and Sensitivity of Avian Magnetoreception", *Physical Review Letters* 109: 110502.

Banaschewski, T., Woerner, W., Rothenberger, A. (2003), "Premonitory sensory phenomena and suppressibility of tics in Tourette syndrome: developmental aspects in children and adolescents", *Developmental Medicine & Child Neurology* 45: pp.700–703.

Barnes, E. (2013), "Emergence and Fundamentality", *Mind* 121: pp.873-901.

Barnes, E.C. (2013), "Freedom, Creativity, and Manipulation", *Noûs*, DOI: 10.1111/nous.12043.

Batterman, R.W. (2002), *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*, New York: Oxford University Press.

---------- (2011), "Emergence, Singularities, and Symmetry Breaking", *Foundations of Physics* 41: pp.1031–1050.

---------- (2014), "Reduction and multiple realizability", paper available in www.robertbatterman.org.

Baumeister, R.F. (2010), "The Self", in Baumeister, R.F, Finkel, E.J., eds., Advanced social psychology: The state of the science, New York: Oxford University Press, pp.139-175.

Baumeister, R.F., Mele, A.R., Vohs, K.D., eds. (2010), *Free Will and Consciousness. How Might They Work?*, Oxford: Oxford University Press.

Bayne, T. (2004), "Self-consciousness and the Unity of Consciousness", *The Monist* 87: pp.224-241.

---------- (2008), "The Unity of Consciousness and the Split-brain Syndrome", *The Journal of Philosophy* 105: pp.277-300.

---------- "The Disunity of Consciousness in Psychiatric Disorders", in Fulford, K.W.M, Davis, M., Gipps, R.G.T., Graham, G., Sadler, J.Z., Stanghellini, G., Thornton, T., eds. (2013), *The Oxford Handbook of Philosophy and Psychiatry*, pp.673-688.

Bedau, M.A., Humphreys, P. (2008), *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge, MA: MIT Press.

Beebee, H., Hitchcock, C., Menzies, P., eds. (2009), *Oxford Handbook of Causation*, Oxford: Oxford University Press.

Bennett, J. (1974), "The Conscience of Huckleberry Finn", *Philosophy* 49, pp. 123–34.

Bennett, K. (2011), "Construction area (no hard hat required)", *Philosophical Studies* 154: pp. 79–104.

Bennett, M.R., Hacker, P. (2003), *Philosophical Foundations of Neuroscience*, Oxford: Blackwell Publishing.

Berofsky, B. (2002), "Ifs, Cans, and Free Will: The Issues." in Kane, R., ed., *The Oxford Handbook of Free Will*, New York: Oxford University Press, pp. 181–201.

Bishop, R.C. (2005), "Patching physics and chemistry together", *Philosophy of Science*72: pp.710–722.

---------- (2006), "The hidden premiss in the causal argument for physicalism", *Analysis* 66: pp.44-52.

---------- (2009), "Whence chemistry? Reductionism and neoreductionism", *Studies in History and Philosophy of Modern Physics* 41: pp.171-177.

---------- (2010), "Free will and the causal closure of physics", in Chiao, R.Y., Cohen, M.L., Leggett, A.J., Phillips, W.D., and Harper, C.L., eds., *Visions of discovery: new light on physics, cosmology, and consciousness*, Cambridge: Cambridge University Press, pp.601-611.

---------- (2011), "Chaos, indeterminism and free will", in Kane, R., ed.,*The Oxford Handbook of Free Will*, 2nd Edition, Oxford: Oxford University Press, pp.84-100.

Bishop, R.C., Atmanspacher, H. (2006), "Contextual emergence in the description of properties", *Foundations of Physics* 36: pp.1753–1777.

Bishop, R.C., Silberstein, M. (n.d.), "Contextual Emergence", unpublished typescript.

Block, N. (1995), "On a confusion about a function of consciousness", *Behavioral and Brain Sciences* 18: pp.227–47.

Bohm, D. (1952), "A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables, I and II", *Physical Review* 85, pp. 166–93.

Bratman, M.E. (1987), *Intention, Plans, and Practical Reason*, Cambridge, MA: Harvard University Press.

---------- (2000), "Reflection, Planning, and Temporally Extended Agency", *Philosophical Review* 109: pp.35-61.

---------- (2005), "Planning Agency, Autonomous Agency", in Taylor, J.S., ed., *Personal Autonomy. New Essays on Personal Autonomy and its Role in Contemporary Moral Philosophy*, New York: Cambridge University Press, pp.35-57.

---------- (2007), "Introduction", in *The Structure of Agency: Essays*, New York: Oxford University Press, pp.3-17.

Brembs, B. (2011), "Toward a scientific concept of free will as a biological trait: spontaneous action and decision-making in invertebrates", *Proceedings of the Royal Society B* 278: pp.930-939.

Broad, C.D. (1925), *The mind and its place in nature*, London: Kegan Paul, Trench, Trubner & Co.

---------- (1952), "Determinism, Indeterminism, and Libertarianism", in *Ethics and the History of Philosophy: Selected Essays*, New York: Humanities Press, pp.195-217.

Buchak, L. (2013), "Free acts and chance: why the rollback argument fails", *The Philosophical Quarterly* 63: pp.20-28.

Buckser, A. (2008), "Before Your Very Eyes: Illness, Agency, and the Management of Tourette Syndrome", Medical Anthropology Quarterly 22: pp.167-192.

Butterfield, H. (1949), *The origins of modern science 1300-1800*, New York: Free Press.

Byrne, A., Hilbert, D.R. (2004), "Hardin, Tye, and Color Physicalism", *The Journal of Philosophy* 101: pp. 37-43.

Campbell, L., Garnett, W., eds. (2010), *The Life of James Clerk Maxwell. With a Selection from his Correspondence and Occasional Writings and a Sketch of his Contributions to Science*, New York: Cambridge University Press.

Cartwright, N. (1999), *The Dappled World: A Study of the Boundaries of Science*, Cambridge: Cambridge University Press.

Chaitin, G. (1975), "Randomness and Mathematical Proof", *Scientific American* 232: pp.47-52.

Chalmers, D. (1996), *The Conscious Mind: in Search of a Fundamental Theory*, New York: Oxford University Press.

---------- (2003), "Consciousness and its Place in Nature", in Stich, S.P., Warfield, T.A., eds., *The Blackwell Guide to Philosophy of Mind*, Oxford: Blackwell Publishing, pp.102-142.

---------- (2010), *The Character of Consciousness*, New York: Oxford University Press.

Chappell, V., ed. (1999), *Hobbes and Bramhall on Liberty and Necessity*, New York: Cambridge University Press.

Chisholm, R. (1964), "J. L. Austin's Philosophical Papers" in Berofsky, B., ed. (1966), *Free Will and Determinism*, New York: Harper & Row: pp.339-45.

---------- (1964), "Human Freedom and the Self" in Kane, R., ed. (2002), *Free Will*, Oxford: Blackwell, pp.47-58.

Chomsky, N. (2000), *New horizons in the study of language and mind*, Cambridge: Cambridge University Press.

Churchland, P.S. (1986), *Neurophilosophy: toward a unified science of the mind/brain*, Cambridge MA: MIT Press.

Clark, P., Butterfield, J. (1987), "Determinism and Probability in Physics", *Proceedings of the Aristotelian Society, Supplementary Volumes* 61: pp.185-243.

Clarke, R. (1995), "Indeterminism and Control", *American Philosophical Quarterly* 32: pp.125-138.

---------- (1996), "Contrastive Rational Explanation of Free Choice", *The Philosophical Quarterly* 46: pp.185-201.

---------- (2003), *Libertarian Accounts of Free Will*, New York: Oxford University Press.

---------- (2008), "Dispositions, Abilities to Act and Free Will: The New Dispositionalism", *Mind* 118: pp.323-51.

---------- (n.d.), "Powers, causes and free will", paper presented at the Freedom and Responsibility Conference at the Queens College, Oxford (March 2014).

Collini, E., Wong, C. Y., Wilk, K. E., Curmi, P. M. G., Brumer, P. & Scholes, G. D. (2010), "Coherently wired light-harvesting in photosynthetic marine algae at ambient temperature", *Nature* 463: pp.644–647.

Corradini, A., O'Connor, T., eds. (2010), *Emergence in Science and Philosophy*, New York: Routledge.

Davidson, D. (1973), "Freedom to act", in Honderich, T., ed., *Essays on freedom and action*, London: Routledge & Kegan Paul, pp.67-86.

---------- (1980), *Essays on Action and Events*, Oxford: Oxford University Press.

---------- (1982), "Rational Animals", *Dialectica* 36: pp.317-327.

Dennett, D.C. (1993), *Consciousness Explained*, London: Penguin (original work: 1991).

---------- (1995), "The Unimagined Preposterousness of Zombies", *Journal of Consciousness Studies* 2: pp.322–6.

---------- (2003), *Freedom Evolves*, London: Penguin.

Descartes, R. (1649), *Les passions de l'âme*, in Adam, C., Tannery, P., eds., (1964–1974), *Oeuvres de Descartes*, vol.XI, Paris: Vrin/CNRS.

---------- (1664), *L'Homme*, in Adam, C., Tannery, P., eds., (1964–1974), *Oeuvres de Descartes*, vol.XI, Paris: Vrin/CNRS.

Dupré, H. (2001), *Human nature and the limits of science*, Oxford: Clarendon Press.

Earman, J. (1971), "Laplacian determinism, or is this any way to run a universe?", *Journal of Philosophy* 68: pp.729-744.

Eccles, J.C. (1970), *Facing Reality. Philosophical Adventures by a Brain Scientist*, Heidelberg: Springer-Verlag.

---------- (1994), *How the SELF Controls Its BRAIN*, Berlin: Springer-Verlag.

---------- (1989), *Evolution of the Brain: Creation of the Self*, New York: Routledge.

Ekstrom, L.W. (2000), *Free Will. A philosophical study*, Boulder, CO: Westview Press.

Ellis, G.F.R. (2009), "Top-down causation and the human brain" in Murphy, N., O'Connor, T., Ellis, G.F.R., eds., *Downward causation and the neurobiology of free will*, Berlin: Springer, pp.63-81.

Engel, G. S. Calhoun, T. R., Read, E. L., Ahn, T.-K., Mancal, T., Cheng, Y.-C., Blankenship, R. E., Fleming, G. R. (2007), "Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems", *Nature* 446: pp.782–786.

Esfeld, M. (2000), "Is Quantum Indeterminism Relevant to Free Will?", *Philosophia Naturalis* 37: pp.177–187.

Everett, H. (1957), "Relative State Formulation of Quantum Mechanics", *Reviews of Modern Physics* 29: pp.454–62.

Fara, M. (2008), "Masked Abilities and Compatibilism", *Mind* 117: pp.844-865.

Fischer, J.M. (1983), "Incompatibilism", *Philosophical Studies* 43: pp.127-37.

---------- (1999), "Frankfurt-style Examples: Responsibility and Semi-compatibilism", reprinted in Kane, R., ed. (2002), *Free Will*, Oxford: Blackwell Publishing, pp.95-109.

---------- (2006), *My Way. Essays on moral responsibility*, Oxford: Oxford University Press.

---------- (2011a), "Frankfurt-type examples and semicompatibilism: new work", in Kane, R., ed., *The Oxford Handbook of Free* Will, 2nd Edition, Oxford: Oxford University Press, pp.243-265.

---------- (2011b), "The Zygote argument remixed", *Analysis* 71: pp.1-6.

Fischer, J.M., Ravizza, M. (2000), *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge: Cambridge University Press.

Frankfurt, H. (1969), "Alternate possibilities and moral responsibility", *Journal of Philosophy* 66: pp.829-839.

---------- (1971), "Freedom of the Will and the Concept of a Person", *Journal of Philosophy* 68: pp.5-20, reprinted in Frankfurt, H. (1988), *The importance of what we care about*, New York: Cambridge University Press, pp.11-25.

Franklin, C. (2011a), "Farewell to the luck (and Mind) argument", *Philosophical Studies* 56: pp.199–230.

---------- (2011b) "Masks, Abilities, and Opportunities: Why the New Dispositionalism Cannot Succeed"*, The Modern Schoolman* (now *Res Philosophica*) 88: pp.89-103.

---------- (2013a) "The Scientific Plausibility of Libertarianism", in Haji, I., Caouette, J., eds., *Free Will and Moral Responsibility*, Newcastle upon Tyne: Cambridge Scholars Publishing, pp. 123-141.

---------- (2013b), "How should libertarians conceive of the location and role of Indeterminism?", *Philosophical Explorations* 16: pp.44-58.

---------- (2014a), "Bratman on Identity over Time and Identification at a Time", paper presented at the APA Central division meeting in Chicago, February 2014.

---------- (2014b), "Event-Causal Libertarianism, Functional Reduction, and the Disappearing Agent Argument", *Philosophical Studies* 170: pp.413-432.

---------- (forthcoming a), "Agent-Causation, Explanation, and *Akrasia*: A Reply to Levy's Hard Luck", *Criminal Law and Philosophy*.

---------- (forthcoming b), "Everyone Thinks that an Ability to Do Otherwise Is Necessary for Free Will and Moral Responsibility", *Philosophical Studies*.

---------- (forthcoming c), "If Anyone Should Be an Agent-Causalist, then Everyone Should Be an Agent-Causalist", *Mind*.

---------- (forthcoming d), "The Luck and Mind arguments", forthcoming in Griffith, M., Levy, N., Timpe, K., eds., *The Routledge Companion to Free Will*, New York: Routledge.

Fried, I., Mukamel, R., Kreiman, G. (2011), "Internally generated preactivation of single neurons in human medial frontal cortex predicts volition", *Neuron* 69: pp.548–562.

Gardner, M. (1970), "Mathematical games. The fantastic combinations of John Conway's new solitaire game 'life'", *Scientific American* 223: pp.120-123.

Gauger, E., Rieper, E., Morton, J.J.L., Benjamin, S.C., Vedral, V. (2011), "Sustained quantum coherence and entanglement in the avian compass", *Physical Review Letters* 106: 040503.

Ghirardi, G.C., Rimini, A., Weber, T. (1986), "Unified Dynamics for Microscopic and Macroscopic Systems", *Physical Review D* 34: pp.470–91.

Gibb, S.C. (2014), "Mental Causation", *Analysis Reviews* 74, pp.327–338.

Ginet, C. (1990), *On Action*, Cambridge: Cambridge University Press.

---------- (1996), "Might We Have No Choice?" in Lehrer, K., ed., *Freedom and Determinism*, New York: Random House, pp.87-104.

Glimcher, P. (2005), "Indeterminacy in Brain and Behavior", *Annual Review of Psychology* 56: pp.25–56.

Gold, J.I., Shadlen, M.N. (2007), "The neural basis of decision making", *Annual Review of Neuroscience* 30: pp.535–74.

Goodman, N. (1965), *Fact, Fiction and Forecast,* New York: Bobbs-Merrill.

Grant, J.E. (2006), "Understanding and Treating Kleptomania: New Models and New Treatments", *Israel Journal of Psychiatry and Related Sciences* 43: pp.81–87.

Grant, J.E., Kim, S. W (2002), "Clinical Characteristics and Associated Psychopathology of 22 Patients with Kleptomania", *Comprehensive Psychiatry* 43: pp.378–384.

Grant, J.E., Odlaug, B. L., Kim, S. W. (2010), "Kleptomania: Clinical Characteristics and Relationship to Substance Use Disorders", *The American Journal of Drug and Alcohol Abuse* 36: pp.291–295.

Griffith, M. (2010), "Why agent-caused actions are not lucky", *American Philosophical Quarterly* 47: pp.43-56.

Hacking, I. (1983), "Nineteenth Century Cracks in the Concept of Determinism", *Journal of the History of Ideas* 44: pp.455-475.

---------- (1990), "Probability and determinism, 1650-1900", in Cantor, G.N., Olby, R.C., Christie, J.R.R., Hodge, M.J.S., eds., *Companion to the History of Modem Science*, London: Routledge, pp.690-701.

Haggard, P. (2008), "Human volition: towards a neuroscience of will", *Nature Reviews Neuroscience* 9: pp.934-946.

Haji, I. (2001), "Control conundrums: modest libertarianism, responsibility, and explanation", *Pacific Philosophical Quaterly* 82: pp.178-200.

---------- (2004), "Active control, agent-causation and free Action", *Philosophical Explorations* 7: pp.131-148.

Hameroff, S., Penrose, R. (2014), "Consciousness in the universe. A review of the 'Orch OR' theory", *Physics of Life Reviews* 11: pp.39–78.

Hanes, D.P., Schall, J.D. (1996), "Neural control of voluntary movement initiation", *Science* 274: pp.427-430.

Hasker, W. (1999), *The emergent self*, Ithaca: Cornell University Press.

Heil, J. (2011), "Powers and the Realization Relation", *The Monist* 94: pp. 34-53.

Heil, J., Mele, A., eds. (1993), *Mental Causation*, Oxford: Clarendon Press.

Hendry, R.F. (2011), "Philosophy of Chemistry", in French, S., Saatsi, J., eds., *The Continuum Companion to the Philosophy of Science*, London: Continuum, pp.293-313.

Hildner, R., Brinks, D., Nieder, J.B., Cogdell, R.J., van Hulst, N.F. (2013), "Quantum coherent energy transfer over varying pathways in single light-harvesting complexes", *Science* 340: pp.1448–1451.

Hill, C.S., McLaughlin, B.P. (1999), "There are Fewer Things in Reality Than Are Dreamt of in Chalmers's Philosophy", *Philosophy and Phenomenological Research* 59: pp.446–454.

Hitchcock, C. (1999), "Contrastive Explanation and the Demons of Determinism", *British Journal for the Philosophy of Science* 50: pp.585–612.

Hobart, R.E. (1934), "Free will as involving determination and inconceivable without it", *Mind* 43: pp.1-17.

Hodgson, D. (1991), *The Mind Matters*, Oxford: Oxford University Press.

---------- (1999), “Hume’s Mistake”, in Libet, B., Freeman, A., Sutherland, K., eds., *The Volitional Brain. Towards a Neuroscience of Free Will*, Exeter: Imprint Academic.

---------- (2005), “A Plain Person’s Free Will”, *Journal of Consciousness Studies* 12: pp.1-19.

---------- (2012), *Rationality + Consciousness = Free Will*, New York: Oxford University Press.

Holton, G., Brush, S. (1985), *Introduction to concepts and theories in physical science*, Princeton: Princeton University Press.

Hooker, C. (2011), *Philosophy of Complex Systems*, Amsterdam: Elsevier.

Honderich, T. (1993), *How free are you?*, Oxford: Oxford University Press.

Horgan, T. (1989), “Mental Quausation”, *Philosophical Perspectives* 3: pp.47-76.

Hornsby, J. (1980), *Actions*, London: Routledge & Kegan Paul.

Hoyt, C.L., Burnette, J.L., Auster-Gussman, L. (2014), “’Obesity Is a Disease’: Examining the Self-Regulatory Impact of This Public-Health Message”, *Psychological Science* 25: pp.997-1002.

Huelga, S.F., Plenio, M.B. (2013), “Vibrations, quanta and biology”, *Contemporary Physics* 54: pp.181-207.

Hume, D. (1748), *Enquiries concerning the human understanding and concerning the principles of morals*, Selby-Bigge, L.A., ed. (1966), Oxford: Clarendon Press.

Humphreys, P. (1997), “How Properties Emerge”, *Philosophy of Science* 64: pp.1-17.

Hunt, D. (2000), “Moral Responsibility and Avoidable Action”, *Philosophical Studies* 97: pp.195-227.

Hütterman, A. (2005), “Explanation, Emergence, and Quantum Entanglement”, *Philosophy of Science* 72: pp.114–127.

Huxley, T. H. (1986), *Lessons in Elementary Physiology*, London: Macmillan (original work: 1866).

Jackson, F. (1982), “Epiphenomenal qualia”, in Lycan, W.G., Prinz, J.J., eds. (1999), *Mind and Cognition. An Anthology*, 2nd Edition, Oxford: Blackwell Publishers, pp.440-446.

---------- (1986), “What Mary didn’t know”, *Journal of Philosophy* 83: pp.291-295.

Jacobs, J.D., O’Connor, T. (2012), "Agent Causation in a Neo-Aristotelian Metaphysics”, in Lowe, EJ, Gibb, S. Ingthorsson, R.D., eds., *Mental Causation and Ontology*, Oxford: Oxford University Press, pp.173-92.

Juarrero, A. (1999), *Dynamics in Action. Intentional Behavior as a Complex System*, Cambridge, MA: MIT Press.

Kane, R. (1998), *The significance of free will*, New York: Oxford University Press (original work: 1996).

---------- (2005), *A contemporary introduction to Free Will*, New York: Oxford University Press.

---------- , ed. (2002), *Free Will*, Oxford: Blackwell Publishing.

Kant, I. (1781), *Critique of Pure Reason*, Guyer, P., Wood, A., eds. (1998), Cambridge: Cambridge University Press.

---------- (1788), *Critique of Practical Reason*, in Gregor, M., trans. and ed. (1996), *Practical Philosophy. The Cambridge Edition of the Works of Immanuel Kant in Translation*, Cambridge: Cambridge University Press.

Kapitan, T. (2002), "A Master Argument for Incompatibilism?", in Kane, R., ed. *The Oxford Handbook of Free Will*, Oxford: Oxford University Press, pp.127-157.

Kennett, J. (2013), "Addiction, Choice and Disease: How Voluntary Is Voluntary Action in Addiction?", in Vincent, N.A., ed., *Neuroscience and Legal Responsibility*, Oxford: Oxford University Press.

Kim, J. (1993), "The Non-Reductivist's Troubles with Mental Causation," in Heil, J., Mele, A., eds., *Mental Causation*, Oxford: Oxford University Press, pp.189-210.

---------- (1993), *Supervenience and Mind: Selected Essays*, Cambridge: Cambridge University Press.

---------- (1999), "Making Sense of Emergence", *Philosophical Studies* 95: pp.3-36.

---------- (2003), "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation", *Philosophy and Phenomenological Research* 67: pp.151-176.

Kircher, T., David, A.S., eds. (2003), *The Self in Neuroscience and Psychiatry*, New York: Cambridge University Press.

Krajbich, I., Armel, C., Rangel, A. (2010), "Visual fixations and the computation and comparison of value in simple choice", *Nature neuroscience* 13: pp.1292-1298.

Kripke, S. (1972), *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Laughlin, R.B., Pines, D. (2000), ‘‘The Theory of Everything’’, *Proceedings of the National Academy of Sciences* 97, reprinted in Bedau, M.A., Humphreys, P. (2008), *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge, MA: MIT Press, pp.259-268.

Lee, H., Cheng, Y. C., Fleming, G. R. (2007), “Coherence dynamics in photosynthesis: protein protection of excitonic coherence”, *Science* 316, pp.1462–1465.

Lehrer, K. (1964), “’Could’ and Determinism”, *Analysis* 24: pp. 159-60.

---------- (1968), “Cans Without Ifs”, *Analysis* 29: pp.29-32.

Leibniz, G.H. (1719), *Theodicy. Essays on the goodness of god, the freedom of man and the origin of evil*, Farrer, A. ed. (1952), New Haven: Yale University Press.

Levine, J. (1983), "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly* 64: pp.354-361.

Levy, N. (2013) “The importance of awareness”, *Australasian Journal of Philosophy* 91: pp.211-229.

---------- (2014), *Consciousness and Moral Responsibility*, Oxford: Oxford University Press.

Lewis, D. (1981), “Are We Free to Break the Laws?”, *Theoria* 47: pp.113-21.

---------- (1986) “Events,” in *Philosophical Papers Volume II*, Oxford: Oxford University Press, pp. 241–69.

Libet, B., Gleason, C.A., Wright E.W., Pearl D.K. (1983), **“**Time of unconscious intention to act in relation to onset of cerebral activity (Readiness-Potential)”, *Brain* 106: pp.623–642.

Lipton, P. (1990), “Contrastive explanation”, in Knowles, D., ed., *Explanation and its limits*, Cambridge: Cambridge University Press, pp.247-66.

---------- (1991), *Inference to the Best Explanation*, London: Routledge.

---------- (1993), “Making a Difference“, *Philosophica* 51: pp.39-54.

Locke, J. (1689), *An essay concerning human understanding*, Nidditch, P.N., ed. (1975), Oxford: Clarendon Press.

Lockwood, M. (2003), “Consciousness and the quantum world: putting qualia in the map”, in Smith, Q., Jokic, A., eds., *Consciousness: New philosophical perspectives*, Oxford: Clarendon Press, pp.447-467.

Lowe, E.J. (2006), “Non-Cartesian Substance Dualism and the Problem of Mental Causation”, *Erkenntnis* 65: pp.5-23.

---------- (2008), *Personal Agency*, New York: Oxford University Press.

London, F., Bauer, E. (1939), *La théorie de l'observation en mécanique quantique*, Paris: Hermann.

Ludwig, K. (1995), “Why the Difference Between Quantum and Classical Physics is Irrelevant to the Mind/Body Problem”, *Psyche* 2(16), available only online at: http://www.theassc.org/files /assc/2350.pdf.

Lycan, W.G., Prinz, J.J., eds. (1999), *Mind and Cognition. An Anthology*, 2nd Edition, Oxford: Blackwell Publishers.

Maier, J. (2014), "Abilities", *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/fall2014/entries/abilities/>.

Margenau, H. (1984), *The Miracle of Existence,* Woodbridge: Ox Bow Press.

McCann, H.J. (1998), *The Works of Agency: On Human Action, Will, and Freedom*, Ithaca, N.Y.: Cornell University Press.

McKay, T., Johnson, D. (1996), "A Reconsideration of an Argument against Compatibilism", *Philosophical Topics* 24: pp.113-122.

McKenna, M. (1997), “Alternative possibilities and the failure of the counter-example strategy”, *Journal of Social Philosophy* 28: pp.71-85.

---------- (2005), “The Relationship Between Autonomous and Morally Responsible Agency”, in Taylor, J.S., ed., *Personal Autonomy*, Cambridge: Cambridge University Press, pp.205-34.

---------- (2008), “A Hard-line Reply to Pereboom’s Four-Case Manipulation Argument”, *Philosophy and Phenomenological Research* 77: pp.142-159.

McKenna, M., Coates, D. J. (2015), "Compatibilism", *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/ archives/sum2015/entries/compatibilism/>.

McLaughlin, B.P. (1992), “The Rise and Fall of British Emergentism”, in Bedau, M.A., Humphreys, P., eds. (2008), *Emergence. Contemporary Readings in Philosophy and Science*, Cambridge, MA: MIT Press, pp.19-59.

Mele, A. (1992), *Springs of Action: Understanding intentional behavior*, New York: Oxford University Press.

---------- (1995), *Autonomous Agents*, New York: Oxford University Press.

---------- (2003), "Agents' Abilities", *Nous* 37: pp.447-470.

---------- (2006), *Free Will and Luck*, New York: Oxford University Press.

---------- (2009), *Effective Intentions. The power of conscious will*, New York: Oxford University Press.

---------- (2013), "Manipulation, Moral Responsibility, and Bullet Biting", *The Journal of Ethics* 17: pp.167-184.

---------- (2014a), *A Dialogue on Free Will and Science*, New York: Oxford University Press.

---------- (2014b), *Free. Why Science Hasn't Disproved Free Will*, New York: Oxford University Press.

Mele, A., Robb, D. (1998), "Rescuing Frankfurt-style Cases", *Philosophical Review* 107: pp. 97-112.

Mercer, I. P. El-Taha, Y. C., Kajumba, N., Marangos, J. P., Tisch, J. W. G., Gabrielsen, M., Cogdell, R. J., Springate, E. Turcu, E. (2009), "Instantaneous mapping of coherently coupled electronic transitions and energy transfers in a photosynthetic complex using angle-resolved coherent optical wave-mixing", *Physical Review Letters* 102, 057402.

Merskey, H., Bogduk, N. (eds.) and IASP Task Force on Taxonomy (1994), *Classification of Chronic Pain*, Second Edition, Seattle: IASP Press.

Mill, J.S. (1868), *A system of logic. Ratiocinative and inductive*, London: Longmans, Green, Reader, and Dyer (original work: 1843).

Miller, W.R., Westerberg, V.S., Harris, R.J., Tonigan, J.S. (1996), "What predicts relapse? Prospective testing of antecedent models", *Addiction* 91: pp.155-172.

Morgan, C.L. (1923), *Emergent Evolution*, London: Williams & Norgate.

Mumford, S., Anjum, R.L. (2011), *Getting Causes from Powers*, Oxford: Oxford University Press.

Murakami, M., Vicente, M.I., Costa, G.M., Mainen, Z.F. (2014), "Neural antecedents of self-initiated actions in secondary motor cortex", *Nature Neuroscience* 17: pp.1574–1582.

Murphy, N., Brown, W. (2007), *Did my neurons make me do it?*: *Philosophical and Neurobiological Perspectives on Moral Responsibility and Free Will*, New York: Oxford University Press.

Murphy, N., O'Connor, T., Ellis, G.F.R., eds. (2009), *Downward causation and the neurobiology of free will*, Berlin: Springer.

Nagel, E. (1961), *The Structure of Science: Problems in the Logic of Scientific Explanation*, New York: Harcourt, Brace, & World.

Nagel, T. (1974), "What is it like to be a bat?", *Philosophical Review* 83: pp.435-450.

---------- (1986), *The View from Nowhere*, New York: Oxford University Press.

---------- (2012), *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False*, New York: Oxford University Press.

Nahmias, E. (2011), "Why 'Willusionism' Leads to 'Bad Results': Comments on Baumeister, Crescioni, and Alquist", *Neuroethics* 4: pp.17-24.

Nahmias, E., Coates, J., Kvaran, T. (2007), "Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions", in *Midwest Studies in Philosophy* 31: pp.214-242.

Nahmias, E., Morris, S.G., Nadelhoffer, T., Turner, J. (2005), "Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility", *Philosophical Psychology* 18: pp.561-84.

---------- (2006) "Is Incompatibilism Intuitive?", *Philosophy and Phenomenological Research* 73: pp.28-53.

Nahmias, E., Murray, D. (2010), "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions", in Aguilar, J., Buckareff, A., Frankish, K., eds. *New Waves in Philosophy of Action*, Palgrave Macmillan, pp.189-215.

Nahmias, E., Shepard, J., Reuter, S. (2014), "It's OK if 'my brain made me do it': People's intuitions about free will and neuroscientific prediction", *Cognition* 133: pp.502-16.

Naylor, M.B. (1984), "Frankfurt on the principle of Alternate possibilities", *Philosophical Studies* 46: pp.249-258.

Newsome, W. (2014), "Neuroscience, Explanation and the Problem of Free Will", in Sinnott-Armstrong, W., ed., *Moral Psychology. Volume 4: Free Will and Moral Responsibility*, Cambridge, MA: MIT Press, pp.81-95.

Noël, X. Van Der Linden, M., Bechara, A. (2006), "The Neurocognitive Mechanisms of Decision-making, Impulse Control, and Loss of Willpower to Resist Drugs**",** *Psychiatry* 3: pp.30–41.

O'Connor, T. (2000), *Persons and Causes*, New York: Oxford University Press.

O'Connor, T., Churchill, J. (2004), "Reasons explanation and agent control: in search of an integrated account", *Philosophical Topics* 32: pp.241-253.

O'Connor, T., Jacobs, J.D. (2003), "Emergent individuals", *The Philosophical Quarterly* 53: pp.540-555.

---------- (2010), "Emergent individuals and the resurrection", *European Journal for Philosophy of Religion* 2, pp.69–88.

O'Connor, T., Wong, H.Y. (2005), "The Metaphysics of Emergence", *Nôus* 39: pp.658–678.

Olson, E.T. (2007), *What Are We? A Study in Personal Ontology*, New York: Oxford University Press.

Otsuka, M. (1998), "Incompatibilism and the avoidability of blame", *Ethics* 108: pp.685-701.

Papineau, D. (2002), *Thinking about Consciousness*, Oxford: Clarendon Press.

Pacherie E. (2007), "The anarchic hand syndrome and utilization behavior: a window onto agentive self-awareness", *Functional Neurology* 22: pp.211-217.

Peirce, C.S. (1892), "The doctrine of necessity examined", *The Monist* 2: pp.321-337.

Penrose, R. (1989), *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*, Oxford: Oxford University Press.

Pereboom, D. (1995), "Determinism *al dente*", *Noûs* 29, pp.21-45.

---------- (2001), *Living Without Free Will*, Cambridge: Cambridge University Press.

---------- (2002), "The explanatory Irrelevance of Alternative Possibilities", in Kane, R., ed., *Free Will*, Oxford: Blackwell Publishing, pp.111-123.

---------- (2003), "Source incompatibilism and alternative possibilities" in Widerker, D., McKenna, M., eds., *Moral Responsibility and Alternative Possibilities*, Aldershot: Ashgate, pp.185-199.

---------- (2004), "Is Our Conception of Agent-Causation Coherent?", *Philosophical Topics* 32: pp.275–86.

---------- (2005), "Defending Hard Incompatibilism", *Midwest Studies in Philosophy* 29: pp.228-247.

---------- (2007), "Hard Incompatibilism", in Pereboom, D., Fischer, J.M., Kane, R., Vargas, M. (2007), *Four Views on Free Will*, Oxford: Wiley-Blackwell, pp.85-125.

---------- (2008), "A hard-line reply to the multiple-case manipulation argument", *Philosophy and Phenomenological Research* 77: pp.160-170.

---------- (2014a), *Free Will, Agency, and Meaning in Life*, Oxford: Oxford University Press.

---------- (2014b), "The disappearing agent objection to event-causal Libertarianism", *Philosophical Studies* 169: pp.59-69.

---------- et al. (2007), *Four Views on Free Will*, Oxford: Wiley-Blackwell.

Popper, K.R. (1950), "Indeterminism in quantum physics and in classical physics", *British Journal for the Philosophy of Science* 1: pp.117-133 (part 1) and 173-195 (part 2).

Popper, K.R., Eccles, J.C. (1977), *The Self and its Brain: an argument for interactionism*, London: Springer.

Pothos, E.M., Busemeyer, J.R. (2013), "Can quantum probability provide a new direction for cognitive modeling?", *Behavioral and brain sciences* 36, pp.255–327.

Rigato, J. (2015), "Reductionism, Agency and Free Will", *Axiomathes* 25: pp.107–116.

Rigato, J., Murakami, M., Mainen, Z. (2015), "Spontaneous decisions and free will: Empirical results and philosophical considerations", in *Cold Spring Harbor Symposia on Quantitative Biology: Cognition, Volume 79,* Cold Spring Harbor Laboratory Press, in press. DOI: 10.1101/sqb.2014.79.024810.

Roitman, J.D., Shadlen, M.N. (2002), "Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task", *The Journal of Neuroscience* 22: pp.9475–9489.

Roskies, A. (2006), "Neuroscientific challenges to free will and responsibility", *Trends in cognitive sciences* 10: pp.419-423.

Ross, P.W., "The Relativity of Color", *Synthese* 123: pp.105–129.

Russell, B. (1913), "On the Notion of Cause", *Proceedings of the Aristotelian Society* 13: pp.1-26.

Russell, P. (1984), "Sorabji and the Dilemma of Determinism", *Analysis* 44: pp.166-172.

Sachs, O. (1995), *An Anthropologist from Mars*, New York: Alfred A. Knopf.

---------- (1998), *The man who mistook his wife for a hat*, New York: Touchstone (original work: 1970).

Santos, G.C. (2015a), "Upward and Downward Causation from a Relational-Horizontal Ontological Perspective", *Axiomathes* 25, pp.23-40.

---------- (2015b), "Ontological Emergence: How is That Possible? Towards a New Relational Ontology", *Foundations of Science,* DOI: 10.1007/s10699-015-9419-x.

Sartanaer, O. (2015), "Synchronic vs. diachronic emergence: a reappraisal", *European Journal for Philosophy of Science* 5, pp.31-54.

Satel, S., Lilienfeld, S.O. (2013), "Addiction and the Brain-Disease Fallacy", *Frontiers in Psychiatry* 4: pp.1-11.

Schroeder, T. (2005), "Moral Responsibility and Tourette Syndrome", *Philosophy and Phenomenological Research* 71: pp.106-123.

Searle, J.R. (1983), *Intentionality. An Essay on the Philosophy of Mind*, Cambridge: Cambridge University Press.

---------- (1987), *Mind, Brains and Science*, Cambridge, MA: Harvard University Press.

---------- (1992), *The Rediscovery of the Mind*, Cambridge, MA: MIT Press.

---------- (2001), *Rationality in Action*, Cambridge, MA: MIT Press.

---------- (2004), *Mind. A Brief Introduction*, New York: Oxford University Press.

Segrè, E. (1980), *From X-rays to quarks. Modern Physicists and Their Discoveries*, Mineola, N.Y.: Dover Publications.

Shadlen, M.N., Gold, J.I. (2004), "The neurophysiology of decision-making as a window on cognition," in Gazzaniga, M. S., ed. *The Cognitive Neurosciences*, *3rd Edition,* Cambridge, MA: MIT Press, pp.1229–1241.

Shepherd, J. (2012) "Free will and consciousness: Experimental studies", *Consciousness and Cognition* 21: pp.915-927.

---------- (2013), "The apparent illusion of conscious deciding", *Philosophical Explorations* 16: pp. 18-30.

---------- (2015), "Consciousness, Free Will and Moral Responsibility: Taking the Folk Seriously", *Philosophical Psychology* 28: pp.929-946.

Silberstein, M. (2001), "Converging on Emergence. Consciousness, Causation and Explanation", *Journal of Consciousness Studies* 8: pp.61-98.

Smart, J.J.C. (1961), "Free-will, praise and blame", *Mind* 70: pp.483-94.

Smilansky, S. (2002), *Free Will and Illusion*, Oxford: Clarendon Press.

Smith, Q. (2003), "Why cognitive sciences cannot ignore quantum mechanics", in Smith, Q., Jokic, A. eds. *Consciousness: New philosophical perspectives*, Oxford: Clarendon Press, pp.409-446.

Soon, C.S., Brass, M., Heinze, H.-J., Haynes, J.-D. (2008), "Unconscious determinants of free decision in the human brain", *Nature Neuroscience* 11: pp.543–545.

Sorabji, R. (1980), *Necessity, Cause and Blame: Perspectives on Aristotle's Philosophy*, Ithaca, N.Y.: Cornell University Press.

Sperry, R.W. (1969), "A modified concept of consciousness", *Psychological review* 76: pp. 532-536.

---------- (1980), "Mind-brain interaction: mentalism, yes; dualism, no", *Neuroscience* 5: pp.195-206.

Spinoza, B. (1677), *Ethics*, in Curley, E., trans. and ed. (1994), *A Spinoza Reader. The* Ethics *and Other Works*, Princeton: Princeton University Press.

---------- (1674), "Letter LXII" in Spinoza, B. (1955), *On the Improvement of the Understanding. The Ethics. Correspondence*, Mineola, N.Y.: Dover.

Stapp, H.P. (1993), *Mind, Matter, and Quantum Mechanics*, Berlin: Springer-Verlag.

---------- (1995), "Why Classical Mechanics Cannot Naturally Accommodate Consciousness But Quantum Mechanics Can", *Psyche* 2(5), available only online at: http://www.theassc.org/files/assc/2345.pdf.

---------- (2006), "Quantum interactive dualism: an alternative to materialism", *Zygon* 51: pp.599-615.

Steward, H. (2008), "Moral Responsibility and the Irrelevance of Physics: Fischer's Semi-Compatibilism versus Anti-Fundamentalism", *Journal of Ethics* 12: pp.129-145.

---------- (2014), *A metaphysics for freedom*, Oxford: Oxford University Press (original work: 2012).

Stich, S.P., Warfield, T.A., eds. (2003), *The Blackwell Guide to Philosophy of Mind*, Oxford: Blackwell Publishing.

Strawson, G. (1994), "The impossibility of Moral Responsibility", *Philosophical Studies* 75: pp.5-24.

Strawson, P.F. (1962), "Freedom and Resentment", *Proceedings of the British Academy* 48: pp.1-25.

Stump, E. (1996), "Libertarian Freedom and the Principle of Alternative Possibilities", in Howard-Snyder, D., Jordan, J. eds., *Faith, Freedom, and Rationality*, Lanham, MD: Rowman and Littlefield, pp.73-88.

---------- (1999), "Dust, Determinism and Frankfurt: A Reply to Goetz", *Faith and Philosophy* 16: pp.413-22.

Todd, P. (2013), "Defending (a modified version of) the Zygote Argument", *Philosophical Studies* 164: pp.189–203.

Tse, P. (2013), *The Neural Basis of Free Will. Criterial Causation*, Cambridge, MA: MIT Press.

Tversky, A., Kahneman, D. (1974), "Judgment under uncertainty: Heuristics and biases", *Science* 185: pp.1124–31.

---------- (1983), "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment", *Psychological Review* 90: pp.293–315.

Twain, M. (1885), *The Adventures of Huckleberry Finn (Tom Sawyer's Comrade)*, New York: Charles L. Webster and Company.

Van Dyke, M. (1982), *An Album of Fluid Motion*, Stanford, CA.: The Parabolic Press.

Van Inwagen, P. (1983), *An essay on free will*, Oxford: Oxford University Press.

---------- (2000), "Free will remains a mystery", *Philosophical Perspectives* 14: pp.1–19.

Velleman, J.D. (1992), 'What Happens When Someone Acts?', *Mind* 101: pp.461-481.

Vihvelin, K. (2004), "Free Will Demystified: A Dispositional Account", Philosophical Topics 32: pp.427-450.

Vision, G. (2011), *Re-Emergence. Locating Conscious Properties in a Material World*, Cambridge, MA: MIT Press.

Volkow, N., Li, T.-K. (2005), "The neuroscience of addiction", *Nature Neuroscience* 8: pp.1429-1420.

von Holst, E., Mittelstaedt, H. (1950), "Das Reafferenzprinzip. Wechselwirkungen zwischen Zentralnervensystem und Peripherie", *Naturwissenschaften* 37: pp.464–476.

Von Neumann, J. (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton: Princeton University Press (original work: 1932).

Waller, R.R. (2014), "The Threat of Effective Intentions to Moral Responsibility in the Zygote Argument", *Philosophia* 42: pp.209–222.

Walter, H. (2001), *Neurophilosophy of Free Will. From Libertarian Illusions to a Concept of Natural Autonomy*, Cambridge, MA: MIT Press.

Walters, Z.B. (2014), "Quantum dynamics of the avian compass", *Physical Review E* 90: 042710 DOI: 10.1103/PhysRevE.90.042710.

Watson, G. (1975), "Free Agency", *Journal of Philosophy* 72: pp.205-20.

---------- (1982), "Free Agency", in Watson, G., *Free Will*, Oxford: Oxford University Press.

Wegner, D. (2002), *The Illusion of Conscious Will*, Cambridge, MA: MIT Press.

Weisskopf, V.F. (1977), "About Liquids", *Transactions of the New York Academy of Sciences* 38: pp.202-218.

Wheeler, J.A., Zurek, W.H. (1983), *Quantum Theory and Measurement*, Princeton: Princeton University Press.

Widerker, D. (1995), "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities", *The Philosophical Review* 104: pp.247-61.

Wigner, E. (1960), "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," in *Communications in Pure and Applied Mathematics* 13, New York: John Wiley & Sons.

---------- (1967), "Remarks on the Mind-Body Question", in Wigner, E.P., *Symmetries and Reflections*, Bloomington, Indiana University Press, pp.171-84 (original work: 1961).

Wilson, J. (2010), "Non-reductive physicalism and degrees of freedom", *The British Journal for the Philosophy of Science* 61: pp.279-311.

Wittgenstein, L (1953), *Philosophical Investigations*, in Anscombe, G.E.M, trans. and ed. (1986), Oxford: Basil Blackwell.

Wolf, S. (1990), *Freedom within Reason*, New York: Oxford University Press.

Wyma, K. (1997), "Moral Responsibility and the Leeway for Action", *American Philosophical Quarterly* 4: pp.57-70.

Zilhão, A. (2005), "The Pertinence of Incontinence", *Principia* 9: pp.193–211.

---------- (2010a), *Pensar com Risco. 25 Lições de Lógica Indutiva*, Lisboa: Imprensa Nacional – Casa da Moeda.

---------- (2010b), *Animal Racional ou Bípede Implume? Um ensaio sobre acção, explicação e racionalidade*, Lisboa: Guerra e Paz.

---------- (2014), "O problema mente-corpo na primeira década do século XXI: visita guiada a pontos-chave da paisagem fisicista", *Kairos* 9: pp.109-137.

---------- (2015), "Free Will and Rationality", *Axiomathes* 25: pp.93-106.